Bayesian shrinkage

Article · December 2012

Source: arXiv

CITATIONS
6

READS
211

4 authors, including:

Anirban Bhattacharya
Duke University
37 PUBLICATIONS **379** CITATIONS
SEE PROFILE

Debdeep Pati
Texas A&M University
42 PUBLICATIONS **268** CITATIONS
SEE PROFILE

David B Dunson
Duke University
468 PUBLICATIONS **11,941** CITATIONS
SEE PROFILE

Bayesian shrinkage

Anirban Bhattacharya, Debdeep Pati, Natesh S. Pillai, David B. Dunson

January 23, 2013

Abstract

Penalized regression methods, such as L_1 regularization, are routinely used in high-dimensional applications, and there is a rich literature on optimality properties under sparsity assumptions. In the Bayesian paradigm, sparsity is routinely induced through two-component mixture priors having a probability mass at zero, but such priors encounter daunting computational problems in high dimensions. This has motivated an amazing variety of continuous shrinkage priors, which can be expressed as global-local scale mixtures of Gaussians, facilitating computation. In sharp contrast to the corresponding frequentist literature, very little is known about the properties of such priors. Focusing on a broad class of shrinkage priors, we provide precise results on prior and posterior concentration. Interestingly, we demonstrate that most commonly used shrinkage priors, including the Bayesian Lasso, are suboptimal in high-dimensional settings. A new class of Dirichlet-Laplace (DL) priors are proposed, which possess optimal concentration and lead to efficient posterior computation exploiting results from normalized random measure theory. Finite sample performance of Dirichlet-Laplace priors relative to alternatives is assessed in simulations.

KEYWORDS: Bayesian; Convergence rate; High dimensional; Lasso; L_1; Penalized regression; Regularization; Shrinkage prior.

1. INTRODUCTION

High-dimensional data have become commonplace in broad application areas, and there is an exponentially increasing literature on statistical and computational methods for big data. In such settings, it is well known that classical methods such as maximum likelihood estimation break down, motivating a rich variety of alternatives based on penalization and thresholding. Most penalization approaches produce a point estimate of a high-dimensional coefficient vector, which has a Bayesian interpretation as corresponding to the mode of a posterior distribution obtained under a shrinkage prior. For example, the wildly popular Lasso/L_1 regularization approach to regression [28] is equivalent to maximum *a posteriori* (MAP) estimation under a Gaussian linear regression model having a double exponential (Laplace) prior on the coefficients. There is a rich theoretical literature justifying the optimality properties of such penalization approaches [19, 20, 25, 29, 33, 34], with fast algorithms [9] and compelling applied results leading to routine use of L_1 regularization in particular.

The overwhelming emphasis in this literature has been on rapidly producing a point estimate with good empirical and theoretical properties. However, in many applications, it is crucial to be able to obtain a realistic characterization of uncertainty in the parameters, in functionals of the parameters and in predictions. Usual frequentist approaches to characterize uncertainty, such as constructing asymptotic confidence regions or using the bootstrap, can break down in high-dimensional settings. For example, in regression when the number of subjects n is much less than the number of predictors p, one cannot naively appeal to asymptotic normality and resampling from the data may not provide an adequate characterization of uncertainty.

Given that most shrinkage estimators correspond to the mode of a Bayesian posterior, it is natural to ask whether we can use the whole posterior distribution to provide a probabilistic measure of uncertainty. Several important questions then arise. Firstly, from a frequentist perspective, we would like to be able to choose a default shrinkage prior that leads to similar optimality properties to those shown for L_1 penalization and other approaches. However, instead of showing that a particular penalty leads to a point estimator having a minimax optimal rate of convergence under

sparsity assumptions, we would like to obtain a (much stronger) result that the entire posterior distribution concentrates at the optimal rate, i.e., the posterior probability assigned to a shrinking neighborhood (proportional to the optimal rate) of the true value of the parameter converges to one. In addition to providing a characterization of uncertainty, taking a Bayesian perspective has distinct advantages in terms of tuning parameter choice, allowing key penalty parameters to be marginalized over the posterior distribution instead of relying on cross-validation. Also, by inducing penalties through shrinkage priors, important new classes of penalties can be discovered that may outperform usual L_q-type choices.

An amazing variety of shrinkage priors have been proposed in the Bayesian literature, with essentially no theoretical justification for the performance of these priors in the high-dimensional settings for which they were designed. [11] and [3] provided conditions on the prior for asymptotic normality of linear regression coefficients allowing the number of predictors p to increase with sample size n, with [11] requiring a very slow rate of growth and [3] assuming $p \leq n$. These results required the prior to be sufficiently flat in a neighborhood of the true parameter value, essentially ruling out shrinkage priors. [2] considered shrinkage priors in providing simple sufficient conditions for posterior consistency in $p \leq n$ settings, while [27] studied finite sample posterior contraction in $p \gg n$ settings.

In studying posterior contraction in high-dimensional settings, it becomes clear that it is critical to obtain tight bounds on prior concentration. This substantial technical hurdle has prevented any previous results (to our knowledge) on posterior concentration in $p \gg n$ settings for shrinkage priors. In fact, prior concentration is critically important not just in studying frequentist optimality properties of Bayesian procedures but for Bayesians in obtaining a better understanding of the behavior of their priors. Without a precise handle on prior concentration, Bayesians face challenges in choosing shrinkage priors and the associated hyperparameters. It becomes an art to use intuition and practical experience to indirectly induce a shrinkage prior, while focusing on Gaussian scale families for computational tractability. Some beautiful classes of priors have been proposed by [2, 5, 13] among others, with [23] showing that essentially all existing shrinkage priors fall within the Gaussian global-local scale mixture family. One of our primary goals is to obtain theory that can

allow evaluation of existing priors and design of novel priors, which are appealing from a Bayesian perspective in allowing incorporation of prior knowledge and from a frequentist perspective in leading to minimax optimality under weak sparsity assumptions.

Shrinkage priors provide a continuous alternative to point mass mixture priors, which include a mass at zero mixed with a continuous density. These priors are highly appealing in allowing separate control of the level of sparsity and the size of the signal coefficients. In a beautiful recent article, [6] showed optimality properties for carefully chosen point mass mixture priors in high-dimensional settings. Unfortunately, such priors lead to daunting computational hurdles in high-dimensions due to the need to explore a 2^p model space. Continuous scale mixtures of Gaussian priors can potentially lead to dramatically more efficient posterior computation.

Focusing on the normal means problem for simplicity in exposition, we provide general theory on prior and posterior concentration under shrinkage priors. One of our main results is that a broad class of Gaussian scale mixture priors, including the Bayesian Lasso [21] and other commonly used choices such as ridge regression, are sub-optimal. We provide insight into the reasons for this sub-optimality and propose a new class of Dirichlet-Laplace (DL) priors, which possess optimal concentration and lead to efficient posterior computation. We show promising initial results for DL priors relative to a variety of competitors.

2. PRELIMINARIES

In studying prior and posterior computation for shrinkage priors, we require some notation and technical concepts. We introduce some of the basic notation here. Technical details in the text are kept to a minimum, and proofs are deferred to a later section.

Given sequences a_n, b_n, we denote $a_n = O(b_n)$ if there exists a global constant C such that $a_n \leq Cb_n$ and $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. For a vector $x \in \mathbb{R}^r$, $\|x\|_2$ denotes its Euclidean norm. We will use Δ^{r-1} to denote the $(r-1)$-dimensional simplex $\{x = (x_1, \ldots, x_r)^{\mathrm{T}} : x_j \geq 0, \sum_{j=1}^r x_j = 1\}$. Further, let Δ_0^{r-1} denote $\{x = (x_1, \ldots, x_{r-1})^{\mathrm{T}} : x_j \geq 0, \sum_{j=1}^{r-1} x_j \leq 1\}$.

For a subset $S \subset \{1, \ldots, n\}$, let $|S|$ denote the cardinality of S and define $\theta_S = (\theta_j : j \in S)$ for a vector $\theta \in \mathbb{R}^n$. Denote supp(θ) to be the *support* of θ, the subset of $\{1, \ldots, n\}$ corresponding

to the non-zero entries of θ. Let $l_0[q; n]$ denote the subset of $\mathbb{R}^n$ given by

$$l_0[q; n] = \{\theta \in \mathbb{R}^n \ : \ \#(1 \leq j \leq n : \theta_j \neq 0) \leq q\}.$$

Clearly, $l_0[q; n]$ consists of q-sparse vectors θ with $|\text{supp}(\theta)| \leq q$.

Let DE(τ) denote a zero mean double-exponential or Laplace distribution with density $f(y) = (2\tau)^{-1}e^{-|y|/\tau}$ for $y \in \mathbb{R}$. Also, we use the following parametrization for the three-parameter generalized inverse Gaussian (giG) distribution: $Y \sim \text{giG}(\lambda, \rho, \chi)$ if $f(y) \propto y^{\lambda-1}e^{-0.5(\rho y+\chi/y)}$ for $y > 0$.

3. CONCENTRATION PROPERTIES OF GLOBAL-LOCAL PRIORS

3.1 **Motivation**

For a high-dimensional vector $\theta \in \mathbb{R}^n$, a natural way to incorporate sparsity in a Bayesian framework is to use point mass mixture priors

$$\theta_j \sim (1 - \pi)\delta_0 + \pi g_\theta, \quad j = 1, \ldots, n, \tag{1}$$

where $\pi = \text{Pr}(\theta_j \neq 0)$, $\mathbb{E}\{|\text{supp}(\theta)| \mid \pi\} = n\pi$ is the prior guess on model size (sparsity level), and g_θ is an absolutely continuous density on $\mathbb{R}$. It is common to place a beta prior on π, leading to a beta-Bernoulli prior on the model size, which conveys an automatic multiplicity adjustment [26]. [6] established that prior (1) with an appropriate beta prior on π and suitable tail conditions on g_θ leads to a frequentist minimax optimal rate of posterior contraction in the normal means setting. We shall revisit the normal means problem in subsection 3.4.

Although point mass mixture priors are intuitively appealing and possess attractive theoretical properties, posterior sampling requires a stochastic search over an enormous space in complicated models where marginal likelihoods are not available analytically, leading to slow mixing and convergence [23]. Computational issues and considerations that many of the θ_js may be small but not exactly zero has motivated a rich literature on continuous shrinkage priors; for some flavor of the

vast literature refer to [2, 5, 13, 14, 21]. [23] noted that essentially all such shrinkage priors can be represented as global-local (GL) mixtures of Gaussians,

$$\theta_j \sim \mathrm{N}(0, \psi_j \tau), \quad \psi_j \sim f, \quad \tau \sim g, \tag{2}$$

where τ controls global shrinkage towards the origin while the local scales $\{\psi_j\}$ allow deviations in the degree of shrinkage. If g puts sufficient mass near zero and f is appropriately chosen, GL priors in (2) can intuitively approximate (1) but through a continuous density concentrated near zero with heavy tails.

GL priors potentially have substantial computational advantages over variable selection priors, since the normal scale mixture representation allows for conjugate updating of θ and ψ in a block. Moreover, a number of frequentist regularization procedures such as ridge, lasso, bridge and elastic net correspond to posterior modes under GL priors with appropriate choices of f and g. For example, one obtains a double-exponential prior corresponding to the popular L_1 or lasso penalty if f has an exponential distribution. However, unlike variable selection priors (1), many aspects of shrinkage priors are poorly understood. For example, even basic properties, such as how the prior concentrates around an arbitrary sparse θ_0, remain to be shown. Hence, subjective Bayesians face difficulties in incorporating prior information regarding sparsity, and frequentists tend to be skeptical due to the lack of theoretical justification.

This skepticism is somewhat warranted, as it is clearly the case that reasonable seeming priors can have poor performance in high-dimensional settings. For example, choosing $\pi = 1/2$ in prior (1) leads to an exponentially small prior probability of 2^{-n} assigned to the null model, so that it becomes literally impossible to override that prior informativeness with the information in the data to pick the null model. However, with a beta prior on π, this problem can be avoided [26]. In the same vein, if one places i.i.d. $\mathrm{N}(0, 1)$ priors on the entries of θ, then the induced prior on $\|\theta\|$ is highly concentrated around $\sqrt{n}$ leading to misleading inferences on θ almost everywhere. These are simple cases, but it is of key importance to assess whether such problems arise for other priors in the GL family and if so, whether improved classes of priors can be found.

There has been a recent awareness of these issues, motivating a basic assessment of the marginal properties of shrinkage priors for a single θ_j. Recent priors such as the horseshoe [5] and generalized double Pareto [2] are carefully formulated to obtain marginals having a high concentration around zero with heavy tails. This is well justified, but as we will see below, such marginal behavior alone is not sufficient; it is necessary to study the joint distribution of θ on $\mathbb{R}^n$. Specifically, we recommend studying the prior concentration $\mathbb{P}(\|\theta - \theta_0\| < t_n)$ where the true parameter θ_0 is assumed to be sparse: $\theta_0 \in l_0[q_n; n]$ with the number of non-zero components $q_n \ll n$ and

$$t_n = n^{\delta/2} \quad \text{with} \quad \delta \in (0, 1). \tag{3}$$

In models where $q_n \ll n$, the prior must place sufficient mass around sparse vectors to allow for good posterior contraction; see subsection 3.4 for further details. Now, as a first illustration, consider the following two extreme scenarios: i.i.d. standard normal priors for the individual components θ_j *vs.* point mass mixture priors given by (1).

Theorem 3.1. *Assume that $\theta_0 \in l_0[q_n; n]$ with $q_n = o(n)$. Then, for i.i.d standard normal priors on θ_j,*

$$\mathbb{P}(\|\theta - \theta_0\|_2 < t_n) \leq e^{-c\,n}. \tag{4}$$

For point mass mixture priors (1) *with $\pi \sim$ Beta$(1, n+1)$ and g_θ being a standard Laplace distribution $g_\theta \equiv$ DE(1),*

$$\mathbb{P}(\|\theta - \theta_0\|_2 < t_n) \geq e^{-c \max\{q_n, \|\theta_0\|_1\}}. \tag{5}$$

Proof. Using $\|\theta\|_2^2 \sim \chi_n^2$, the claim made in (4) follows from an application of Anderson's inequality (6.1) and standard chi-square deviation inequalities. In particular, the exponentially small concentration also holds for $\mathbb{P}(\|\theta_0\|_2 < t_n)$. The second claim (5) follows from results in [6]. □

As seen from Theorem 3.1, the point mass mixture priors have much improved concentra-

tion around sparse vectors, as compared to the i.i.d. normal prior distributions. The theoretical properties enjoyed by the point mass mixture priors can mostly be attributed to this improved concentration. The above comparison suggests that it is of merit to evaluate a shrinkage prior in high dimensional models under sparsity assumption by obtaining its concentration rates around sparse vectors. In this paper, we carry out this program for a wide class of shrinkage priors. Our analysis also suggests some novel priors with improved concentration around sparse vectors.

In order to communicate our main results to a wide audience, we will first present specific corollaries of our main results applied to various existing shrinkage priors. The main results are given in Section 6. Recall the GL priors presented in (2) and the sequence t_n in (3).

3.2 Prior concentration for global priors

This simplified setting involves only a global parameter, *i.e.*, $\psi_j = 1$ for all j. This subclass includes the important example of ridge regression, with τ routinely assigned an inverse-gamma prior, $\tau \sim \text{IG}(\alpha, \beta)$.

Theorem 3.2. *Assume $\theta \sim \text{GL}$ with $\psi_j = 1$ for all j. If the prior f on the global parameter τ has an $\text{IG}(\alpha, \beta)$ distribution, then*

$$\mathbb{P}(\|\theta\|_2 < t_n) \leq e^{-Cn^{1-\delta}}, \tag{6}$$

where $C > 0$ is a constant depending only on α and β.

The above theorem shows that compared to i.i.d. normal priors (4), the prior concentration does not improve much under an inverse-gamma prior on the global variance regardless of the hyperparameters (provided they don't scale with n) even when $\theta_0 = 0$. Concentration around θ_0 away from zero will clearly be even worse. Hence, such a prior is not well-suited in high-dimensional settings, confirming empirical observations documented in [10, 24]. It is also immediate that the same concentration bound in (6) would be obtained for the giG family of priors on τ.

In [24], the authors instead recommended a half-Cauchy prior as a default choice for the global variance (also see [10]). We consider the following general class of densities on $(0, \infty)$ for τ, to

be denoted $\mathcal{F}$ henceforth, that satisfy: (i) $f(\tau) \leq M$ for all $\tau \in (0, \infty)$ (ii) $f(\tau) > 1/M$ for all $\tau \in (0, 1)$, for some constant $M > 0$. Clearly, $\mathcal{F}$ contains the half-Cauchy and exponential families. The following result provides concentration bounds for these priors.

Theorem 3.3. *Let $\|\theta_0\|_2 = o(\sqrt{n})$. If the prior f on the global parameter τ belongs to the class $\mathcal{F}$ above then,*

$$C_1 e^{-(1-\delta)\log n} \leq \mathbb{P}(\|\theta\|_2 < t_n) \leq C_2 e^{-(1-\delta)\log n}. \tag{7}$$

Furthermore, if $\|\theta_0\|_2 > t_n$, then

$$e^{-c_1 n \log a_n} \leq \mathbb{P}(\|\theta - \theta_0\|_2 < t_n) \leq e^{-c_2 n \log a_n}, \tag{8}$$

where $a_n = \|\theta_0\|_2 / t_n > 1$ and $c_i, C_i > 0$ are constants with C_1, C_2, c_2 depending only on M in the definition of $\mathcal{F}$ and c_1 depending on M and δ.

Thus (7) in Theorem 3.3 shows that the prior concentration around zero can be dramatically improved from exponential to polynomial with a careful prior on τ that can assign sufficient mass near zero, such as the half-Cauchy prior [10, 24]. Unfortunately, as (8) shows, for signals of large magnitude one again obtains an exponentially decaying probability. Hence, Theorem 3.3 conclusively shows that global shrinkage priors are simply not flexible enough for high-dimensional problems.

Remark 3.4. *The condition $\|\theta_0\|_2 \geq t_n$ is only used to prove the lower bound in* (8)*. For any $\|\theta_0\|$ bounded below by a constant, we would still obtain an upper bound $e^{-Cn^{1-\delta}\log n}$ in* (8)*, similar to the bound in* (6)*.*

3.3 Prior concentration for a class of GL priors

Proving concentration results for the GL family (2) in the general setting presents a much harder challenge compared to Theorem 3.3 since we now have to additionally integrate over the n local parameters $\psi = (\psi_1, \ldots, \psi_n)$. We focus on an important sub-class in Theorem 6.4 below, namely

the exponential family for the distribution of g in (2). For analytical tractability, we additionally assume that θ_0 has only one non-zero entry. The interest in the exponential family arises from the fact that normal-exponential scale mixtures give rise to the double-exponential family [32]: $\theta \mid \psi \sim N(0, \psi\sigma^2), \psi \sim \text{Exp}(1/2)$ implies $\theta \sim \text{DE}(\sigma)$, and hence this family of priors can be considered as a Bayesian version of the lasso [21]. We now state a concentration result for this class noting that a general version of Theorem 3.5 can be found in Theorem 6.4 stated in Section 6.

Theorem 3.5. *Assume $\theta \sim \text{GL}$ with $f \in \mathcal{F}$ and $g \equiv \text{Exp}(\lambda)$ for some constant $\lambda > 0$. Also assume θ_0 has only one non-zero entry and $\|\theta_0\|_2^2 > \log n$. Then, for a global constant $C > 0$ depending only on M in the definition of $\mathcal{F}$,*

$$\mathbb{P}(\|\theta - \theta_0\|_2 < t_n) \leq e^{-C\sqrt{n}}. \tag{9}$$

Theorem 3.5 asserts that even in the simplest deviation from the null model with only one signal, one continues to have exponentially small concentration under an exponential prior on the local scales. From (5) in Theorem 3.1, appropriate point mass mixture priors (1) would have $\mathbb{P}(\|\theta - \theta_0\|_2 < t_n) \geq e^{-C\|\theta_0\|_1}$ under the same conditions as above, clearly showing that the wide difference in concentration still persists.

3.4 Posterior lower bounds in normal means

We have discussed the prior concentration for a high-dimensional vector θ without alluding to any specific model so far. In this section we show how prior concentration impacts posterior inference for the widely studied normal means problem [1] (see [6, 8, 15] and references therein):

$$y_i = \theta_i + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, 1), \quad 1 \leq i \leq n. \tag{10}$$

The minimax rate s_n for the above model is given by $s_n^2 = q_n \log(n/q_n)$ when $\theta_0 \in l_0[q_n; n]$.

[1] Although we study the normal means problem, the ideas and results in this section are applicable to other models such as non-parametric regression and factor models.

For this model [6] recently established that for point mass priors for θ with $\pi \sim \text{beta}(1, \kappa n + 1)$ and g_θ having Laplace like or heavier tails, the posterior contracts at the minimax rate, *i.e.,* $\mathbb{E}_{n,\theta_0}\mathbb{P}(\|\theta - \theta_0\|_2 < M s_n \mid y) \to 1$ for some constant $M > 0$. Thus we see that carefully chosen point mass priors are indeed optimal[2]. However not all choices for g_θ lead to optimal proceedures; [6] also showed that if g_θ is instead chosen to be standard Gaussian, *the posterior does not contract at the minimax rate*, *i.e.*, one could have $\mathbb{E}_{n,\theta_0}\mathbb{P}(\|\theta - \theta_0\|_2 < s_n \mid y) \to 0$ for signals of sufficiently large magnitude. This result is particularly striking given the routine choice of Gaussian for g_θ in Bayesian variable selection and thus clearly illustrates the need for careful prior choice in high dimensions.

To establish such a posterior lower-bound result, [6] showed that given a fixed sequence t_n, if there exists a sequence r_n ($r_n > t_n$) such that

$$\frac{\mathbb{P}(\|\theta - \theta_0\|_2 < t_n)}{\mathbb{P}(\|\theta - \theta_0\|_2 < r_n)} = o(e^{-r_n^2}), \tag{11}$$

then $\mathbb{P}(\|\theta - \theta_0\|_2 < t_n \mid y) \to 0$. This immediately shows the importance of studying the prior concentration. Intuitively, (11) would be satisfied when the prior mass of the bigger ball $\|\theta - \theta_0\|_2 < r_n$ is almost entirely contained in the annulus with inner radius t_n and outer radius r_n, so that the smaller ball $\|\theta - \theta_0\|_2 < t_n$ barely has any prior mass compared to the bigger ball. As an illustrative example, in the i.i.d. $\mathrm{N}(0, 1)$ example with $t_n = s_n$, setting $r_n = \sqrt{n}$ would satisfy (11) above, proving that i.i.d. $\mathrm{N}(0, 1)$ priors are sub-optimal. Our goal is to investigate whether a similar phenomenon persists for global-local priors in light of the concentration bounds developed in Theorems 3.3 and 6.4.

As in Section 3.2, we first state our posterior lower bound result for the case where there is only a global parameter.

Theorem 3.6. *Suppose we observe* $y \sim N_n(\theta_0, \mathrm{I}_n)$ *and* (10) *is fitted with a* GL *prior on* θ *such that* $\psi_j = 1$ *for all* j *and the prior* f *on the global parameter* τ *lies in* $\mathcal{F}$. *Assume* $\theta_0 \in l_0[q_n; n]$ *where* $q_n/n \to 0$ *and* $\|\theta_0\|_2 > s_n$, *with* $s_n^2 = q_n \log(n/q_n)$ *being the minimax squared error loss*

[2]It is important that the hyper parameter for π depends on n. We do not know if the result holds without this

over $l_0[q_n; n]$. *Then,* $\mathbb{E}_{n,\theta_0}\mathbb{P}(\|\theta - \theta_0\|_2 \leq s_n \mid y) \to 0$.

Proof. Without loss of generality, assume $\|\theta_0\|_2 = o(\sqrt{n})$, since the posterior mass with a prior centered at the origin would be smaller otherwise. Choosing $t_n = s_n$, r_n to be a sequence such that $t_n < r_n < \|\theta_0\|_2$ and resorting to the two-sided bounds in Theorem 3.3, the ratio in (11) is smaller than $(t_n/r_n)^n$, and hence $e^{r_n^2}(t_n/r_n)^n \to 0$ since $r_n \leq \|\theta_0\|_2 = o(\sqrt{n})$. □

Theorem 3.6 states that a GL prior with only a global scale is sub-optimal if $\|\theta_0\|_2 > s_n$. Observe that in the complementary region $\{\|\theta_0\|_2 \leq s_n\}$, the estimator $\hat{\theta} \equiv 0$ attains squared error in the order of $q_n \log(n/q_n)$, implying the condition $\|\theta_0\|_2 > s_n$ is hardly stringent.

Next, we state a result for the sub-class of GL priors as in Theorem 6.4, i.e., when g has an exponential distribution leading to a double-exponential distribution marginally.

Theorem 3.7. *Suppose we observe* $y \sim N_n(\theta_0, \mathrm{I}_n)$ *and the model in* (10) *is fitted with a* GL *prior on* θ *such that* f *lies in* $\mathcal{F}$ *and* $g \equiv \mathrm{Exp}(\lambda)$ *for some constant* $\lambda > 0$. *Assume* $\theta_0 \in l_0[q_n; n]$ *with* $q_n = 1$ *and* $\|\theta_0\|_2^2 / \log n \to \infty$. *Then,* $\mathbb{E}_{n,\theta_0}\mathbb{P}(\|\theta - \theta_0\|_2 \leq \sqrt{\log n} \mid y) \to 0$.

A proof of Theorem 3.7 is deferred to Section 6. From [6], appropriate point mass mixture priors would assign increasing mass with n to the same neighborhood in Theorem 3.7. Hence, many of the shrinkage priors used in practice are sub-optimal in high-dimensional applications, even in the simplest deviation from the null model with only one moderately sized signal. Although Theorem 3.7 is stated and proved for g having an exponential distribution (which includes the Bayesian lasso [21]), we conjecture that the conclusions would continue to be valid if one only assumes g to have exponential tails plus some mild conditions on the behavior near zero. However, the assumptions of Theorem 3.7 precludes the case when g has polynomial tails, such as the horseshoe [5] and generalized double Pareto [2]. One no longer obtains tight bounds on the prior concentration for g having polynomial tails using the current techniques and it becomes substantially complicated to study the posterior.

Another important question beyond the scope of the current paper should concern the behavior of the posterior when one plugs in an empirical Bayes estimator of the global parameter τ. How-

ever, we show below that the "optimal" sample-size dependent plug-in choice $\tau_n = c^2/\log n$ (so that marginally $\theta_j \sim \text{DE}(c/\sqrt{\log n})$) for the lasso estimator [20] produces a sub-optimal posterior:

Theorem 3.8. *Suppose we observe $y \sim N_n(\theta_0, \mathrm{I}_n)$ and (10) is fitted with a GL prior on θ such that τ is deterministically chosen to be τ_n, i.e., $f \equiv \delta_{\tau_n}$ for a non-random sequence τ_n and $g \equiv \text{Exp}(\lambda)$ for some constant $\lambda > 0$. Assume $\theta_0 \in l_0[q_n; n]$ with $q_n(\log n)^2 = o(n)$ and $\tau_n = c/\log n$ is used as the plug-in choice. Then, $\mathbb{E}_{n,\theta_0}\mathbb{P}(\|\theta - \theta_0\|_2 \leq s_n \mid y) \to 0$, with $s_n^2 = q_n \log(n/q_n)$ being the minimax squared error loss over $l_0[q_n; n]$.*

A proof of Theorem 3.8 can be found in Section 6. Note that a slightly stronger assumption on the sparsity allows us to completely obviate any condition on θ_0 in this case. Also, the result can be generalized to any τ_n if $q_n \log n/\tau_n = o(n)$.

4. A NEW CLASS OF SHRINKAGE PRIORS

The results in Section 3 necessitate the development of a general class of continuous shrinkage priors with improved concentration around sparse vectors. To that end, let us revisit the global-local specification (2). After integrating out the local scales ψ_j's, (2) can be equivalently represented as a global scale mixture of a kernel $\mathcal{K}(\cdot)$,

$$\theta_j \overset{\text{i.i.d.}}{\sim} \mathcal{K}(\cdot\,, \tau), \quad \tau \sim g, \tag{12}$$

where $\mathcal{K}(x) = \int \psi^{-1/2}\phi(x/\sqrt{\psi})g(\psi)d\psi$ is a symmetric unimodal density (or kernel) on $\mathbb{R}$ and $\mathcal{K}(x, \tau) = \tau^{-1/2}\mathcal{K}(x/\sqrt{\tau})$. For example, $\psi_j \sim \text{Exp}(1/2)$ corresponds to a double exponential kernel $\mathcal{K} \equiv \text{DE}(1)$, while $\psi_j \sim \text{IG}(1/2, 1/2)$ results in a standard Cauchy kernel $\mathcal{K} \equiv \text{Ca}(0, 1)$. These traditional choices lead to a kernel which is *bounded* in a neighborhood of zero, and the resulting global-local procedure (12) with a single global parameter τ doesn't attain the desired concentration around sparse vectors as documented in Theorem 3.5, leading to sub-optimal behavior of the posterior in Theorem 3.7.

However, if one instead uses a half Cauchy prior $\psi_j^{1/2} \sim \text{Ca}_+(0, 1)$, then the resulting horseshoe kernel [4, 5] is unbounded with a singularity at zero. This phenomenon coupled with tail

robustness properties leads to excellent empirical performances of the horseshoe. However, the joint distribution of θ under a horseshoe prior is understudied. One can imagine that it achieves a higher prior concentration around sparse vectors compared to common shrinkage priors since the singularity at zero potentially allows most of the entries to be concentrated around zero with the heavy tails ensuring concentration around the relatively small number of signals. However, the polynomial tails of ψ_j present a hindrance in obtaining tight bounds using our techniques. We hope to address the polynomial tails case in details elsewhere, though based on strong empirical performance, we conjecture that the horseshoe leads to the optimal posterior contraction in a much broader domain compared to the Bayesian lasso and other common shrinkage priors. The normal-gamma scale mixtures [13] and the generalized double Pareto prior [2] follow the same philosophy and should have similar properties.

The above class of priors rely on obtaining a suitable kernel $\mathcal{K}$ through appropriate normal scale mixtures. In this article, we offer a fundamentally different class of shrinkage priors that alleviate the requirements on the kernel, while having attractive theoretical properties. In particular, our proposed class of kernel-Dirichlet (kD) priors replaces the single global scale τ in (12) by a vector of scales $(\phi_1\tau, \ldots, \phi_n\tau)$, where $\phi = (\phi_1, \ldots, \phi_n)$ is constrained to lie in the $(n-1)$ dimensional simplex $\mathcal{S}^{n-1}$:

$$\theta_j \mid \phi_j, \tau \sim \mathcal{K}(\cdot\,, \phi_j\tau), \quad (\phi, \tau) \in \mathcal{S}^{n-1} \otimes \mathbb{R}^+, \tag{13}$$

where $\mathcal{K}$ is any symmetric (about zero) unimodal density that can be represented as scale mixture of normals [32]. While previous shrinkage priors in the literature obtain marginal behavior similar to the point mass mixture priors (1), our construction aims at resembling the *joint distribution* of θ under a two-component mixture prior. Constraining ϕ on $\mathcal{S}^{n-1}$ restrains the "degrees of freedom" of the ϕ_j's, offering better control on the number of dominant entries in θ. In particular, letting $\phi \sim \text{Dir}(a, \ldots, a)$ for a suitably chosen a allows (13) to behave like (1) jointly, forcing a large subset of $(\theta_1, \ldots, \theta_n)$ to be *simultaneously* close to zero with high probability.

We focus on the Laplace kernel from now on for concreteness, noting that all the results stated

below can be generalized to other choices. The corresponding hierarchical prior

$$\theta_j \sim \mathrm{DE}(\phi_j \tau), \quad \phi \sim \mathrm{Dir}(a, \ldots, a), \quad \tau \sim g \tag{14}$$

is referred to as a Dirichlet-Laplace prior, denoted $\mathrm{DL}_a(\tau)$. In the following Theorem 4.1, we establish the improved prior concentration of the DL prior. For sake of comparison with the global-local priors in Section 3.3, we assume the same conditions as in Theorem 3.5; a general version can be found in Section 6.

Theorem 4.1. *Assume $\theta \sim DL_a(\tau)$ as in* (14) *with $a = 1/n$ and $\tau \sim \mathrm{Exp}(\lambda)$ for some $\lambda > 0$. Also assume θ_0 has only one non-zero entry and $\|\theta_0\|_2^2 = c \log n$. Also, recall the sequence t_n in* (3)*. Then, for a constant C depending only on δ on λ,*

$$P(\|\theta - \theta_0\| < t_n) \geq \exp\{-C\sqrt{\log n}\}. \tag{15}$$

From (5) in Theorem 3.1, appropriate point mass mixtures would attain exactly the same concentration as in (15), showing the huge improvement in concentration compared to global-local priors. This further establishes the role of the dependent scales ϕ, since in absence of ϕ, a $\mathrm{DE}(\tau)$ prior with $\tau \sim \mathrm{Exp}(\lambda)$ would lead to a concentration smaller than $e^{-C\sqrt{n}}$ (see Theorem 3.5).

To further understand the role of ϕ, we undertake a study of the marginal properties of θ_j integrating out ϕ_j. Clearly, the marginal distribution of ϕ_j is $\mathrm{Beta}(a, (n-1)a)$. Let $\mathrm{WG}(\alpha, \beta)$ denote a wrapped gamma distribution with density function

$$f(x; \alpha, \beta) \propto |x|^{\alpha - 1} e^{-\beta |x|}, \quad x \in \mathbb{R}.$$

The results are summarized in Proposition 4.2 below.

Proposition 4.2. *If $\theta \mid \phi, \tau \sim \mathrm{DL}_a(\tau)$ and $\phi \sim \mathrm{Dir}(a, \ldots, a)$, then the marginal distribution of θ_j given τ is unbounded with a singularity at zero for any $a < 1$. Further, in the special case $a = 1/n$, the marginal distribution is a wrapped Gamma distribution $\mathrm{WG}(1/n, \tau^{-1})$.*

Thus, marginalizing over ϕ, we obtain an unbounded kernel $\mathcal{K}$ (similar to the horseshoe). Since the marginal density of $\theta_j \mid \tau$ has a singularity at 0, it assigns a huge mass at zero while retaining exponential tails, which partly explains the improved concentration. A proof of Proposition 4.2 can be found in the appendix.

There is a recent frequentist literature on including a local penalty specific to each coefficient. The adaptive Lasso [31, 35] relies on empirically estimated weights that are plugged in. [18] instead propose to sample the penalty parameters from a posterior, with a sparse point estimate obtained for each draw. These approaches do not produce a full posterior distribution but focus on sparse point estimates.

4.1 **Posterior computation**

The proposed class of DL priors leads to straightforward posterior computation via an efficient data augmented Gibbs sampler. Note that the $\text{DL}_a(\tau)$ prior (14) can be equivalently represented as

$$\theta_j \sim \text{N}(0, \psi_j \phi_j^2 \tau^2),\ \psi_j \sim \text{Exp}(1/2),\ \phi \sim \text{Dir}(a, \ldots, a).$$

In the general $\text{DL}_a(\tau)$ setting, we assume a gamma$(\lambda, 1/2)$ prior on τ with $\lambda = na$. In the special case when $a = 1/n$, the prior on τ reduces to an Exp$(1/2)$ prior consistent with the statement of Theorem 4.1.

We detail the steps in the normal means setting but the algorithm is trivially modified to accommodate normal linear regression, robust regression with heavy tailed residuals, probit models, logistic regression, factor models and other hierarchical Gaussian cases. To reduce auto-correlation, we rely on marginalization and blocking as much as possible. Our sampler cycles through (i) $\theta \mid \psi, \phi, \tau, y$, (ii) $\psi \mid \phi, \tau, \theta$, (iii) $\tau \mid \phi, \theta$ and (iv) $\phi \mid \theta$. We use the fact that the joint posterior of (ψ, ϕ, τ) is conditionally independent of y given θ. Steps (ii) - (iv) together gives us a draw from the conditional distribution of $(\psi, \phi, \tau) \mid \theta$, since

$$[\psi, \phi, \tau \mid \theta] = [\psi \mid \phi, \tau, \theta][\tau \mid \phi, \theta][\phi \mid \theta].$$

Steps (i) – (iii) are standard and hence not derived. Step (iv) is non-trivial and we develop an efficient sampling algorithm for jointly sampling ϕ. Usual one at a time updates of a Dirichlet vector leads to tremendously slow mixing and convergence, and hence the joint update in Theorem 4.3 is an important feature of our proposed prior.

Theorem 4.3. *The joint posterior of $\phi \mid \theta$ has the same distribution as $(T_1/T, \ldots, T_n/T)$, where T_j are independently distributed according to a giG$(a-1, 1, 2|\theta_j|)$ distribution, and $T = \sum_{j=1}^n T_j$.*

Proof. Integrating out τ, the joint posterior of $\phi \mid \theta$ has the form

$$\pi(\phi_1, \ldots, \phi_{n-1} \mid \theta) \propto \prod_{j=1}^n \left[\phi_j^{a-1} \frac{1}{\phi_j}\right] \int_{\tau=0}^{\infty} e^{-\tau/2} \tau^{\lambda-n-1} e^{-\sum_{j=1}^n |\theta_j|/(\phi_j \tau)} d\tau. \tag{16}$$

We now state a result from the theory of normalized random measures (see, for example, (36) in [17]). Suppose $T_1, \ldots, T_n$ are independent random variables with T_j having a density f_j on $(0, \infty)$. Let $\phi_j = T_j/T$ with $T = \sum_{j=1}^n T_j$. Then, the joint density f of $(\phi_1, \ldots, \phi_{n-1})$ supported on the simplex $\mathcal{S}^{n-1}$ has the form

$$f(\phi_1, \ldots, \phi_{n-1}) = \int_{t=0}^{\infty} t^{n-1} \prod_{j=1}^n f_j(\phi_j t) dt, \tag{17}$$

where $\phi_n = 1 - \sum_{j=1}^{n-1} \phi_j$. Setting $f_j(x) \propto \frac{1}{x^\delta} e^{-|\theta_j|/x} e^{-x/2}$ in (17), we get

$$f(\phi_1, \ldots, \phi_{n-1}) = \left[\prod_{j=1}^n \frac{1}{\phi_j^\delta}\right] \int_{t=0}^{\infty} e^{-t/2} t^{n-1-n\delta} e^{-\sum_{j=1}^n |\theta_j|/(\phi_j t)} dt. \tag{18}$$

We aim to equate the expression in (18) with the expression in (16). Comparing the exponent of ϕ_j gives us $\delta = 2 - a$. The other requirement $n - 1 - n\delta = \lambda - n - 1$ is also satisfied, since $\lambda = na$. The proof is completed by observing that f_j corresponds to a giG$(a - 1, 1, 2|\theta_j|)$ when $\delta = 2 - a$. □

The summary of each step are finally provided below.

(i) To sample $\theta \mid \psi, \phi, \tau, y$, draw θ_j independently from a $\mathrm{N}(\mu_j, \sigma_j^2)$ distribution with

$$\sigma_j^2 = \{1 + 1/(\psi_j \phi_j^2 \tau^2)\}^{-1}, \quad \mu_j = \{1 + 1/(\psi_j \phi_j^2 \tau^2)\}^{-1} y.$$

(ii) The conditional posterior of $\psi \mid \phi, \tau, \theta$ can be sampled efficiently in a block by independently sampling $\psi_j \mid \phi, \theta$ from an inverse-Gaussian distribution $\mathrm{iG}(\mu_j, \lambda)$ with $\mu_j = \phi_j \tau / |\theta_j|, \lambda = 1$.

(iii) Sample the conditional posterior of $\tau \mid \phi, \theta$ from a $\mathrm{giG}(\lambda - n, 1, 2\sum_{j=1}^n |\theta_j|/\phi_j)$ distribution.

(iv) To sample $\phi \mid \theta$, draw $T_1, \ldots, T_n$ independently with $T_j \sim \mathrm{giG}(a - 1, 1, 2|\theta_j|)$ and set $\phi_j = T_j / T$ with $T = \sum_{j=1}^n T_j$.

5. SIMULATION STUDY

Since the concentration results presented here are non-asymptotic in nature, we expect the theoretical findings to be reflected in finite-sample performance. In particular, we aim to study whether the improved concentration of the proposed Dirichlet-Laplace ($\mathrm{DL}_{1/n}$) priors compared to the Bayesian lasso (BL) translate empirically. As illustration, we show the results from a replicated simulation study with various dimensionality n and sparsity level q_n. In each setting, we have 100 replicates of a n-dimensional vector y sampled from a $\mathrm{N}_n(\theta_0, \mathrm{I}_n)$ distribution with θ_0 having q_n non-zero entries which are all set to be a constant $A > 0$. We chose two values of n, namely $n = 100, 200$. For each n, we let $q_n = 5, 10, 20\%$ of n and choose $A = 7, 8$. This results in 12 simulation settings in total. The simulations were designed to mimic the setting in Section 3 where θ_0 is sparse with a few moderate-sized coefficients.

The squared error loss corresponding to the posterior median averaged across simulation replicates is provided in Table 1. To offer further grounds for comparison, we have also tabulated the results for Lasso (LS), Empirical Bayes median (EBMed) as in [15] [3], posterior median with a point mass prior (PM) as in [6] and the posterior median corresponding to the horseshoe prior [5].

[3]The EBMed procedure was implemented using the package [16].

Table 1: Squared error comparison over 100 replicates. Average squared error across replicates reported for BL (Bayesian lasso), DL (Dirichlet-Laplace), LS (Lasso), EBMed (Empirical Bayes median), PM (Point mass prior) and HS (horseshoe).

n	**100**						**200**					
$\frac{\mathbf{q_n}}{\mathbf{n}}\%$	**5**		**10**		**20**		**5**		**10**		**20**	
A	**7**	**8**	**7**	**8**	**7**	**8**	**7**	**8**	**7**	**8**	**7**	**8**
BL	33.05	33.63	49.85	50.04	68.35	68.54	64.78	69.34	99.50	103.15	133.17	136.83
$DL_{1/n}$	8.20	7.19	17.29	15.35	32.00	29.40	16.07	14.28	33.00	30.80	65.53	59.61
LS	21.25	19.09	38.68	37.25	68.97	69.05	41.82	41.18	75.55	75.12	137.21	136.25
EBMed	13.64	12.47	29.73	27.96	60.52	60.22	26.10	25.52	57.19	56.05	119.41	119.35
PM	12.15	10.98	25.99	24.59	51.36	50.98	22.99	22.26	49.42	48.42	101.54	101.62
HS	8.30	7.93	18.39	16.27	37.25	35.18	15.80	15.09	35.61	33.58	72.15	70.23

For the fully Bayesian analysis using point mass mixture priors, we use a complexity prior on the subset-size, $\pi_n(s) \propto \exp\{-\kappa s \log(2n/s)\}$ with $\kappa = 0.1$ and independent standard Laplace priors for the non-zero entries as in [6]. [4] [5]

Even in this succinct summary of the results, a wide difference between the Bayesian Lasso and the proposed $\text{DL}_{1/n}$ is observed in Table 1, vindicating our theoretical results. The horseshoe performs similarly as the $\text{DL}_{1/n}$. The superior performance of the $\text{DL}_{1/n}$ prior can be attributed to its strong concentration around the origin. However, in cases where there are several relatively small signals, the $\text{DL}_{1/n}$ prior can shrink all of them towards zero. In such settings, depending on the practitioner's utility function, the singularity at zero can be "softened" using a DL_a prior for a smaller value of a. Based on empirical performance and computational efficiency, we recommend $a = 1/2$ as a robust default choice. The computational gain arises from the fact that in this case, the distribution of T_j in (iv) turns out to be inverse-Gaussian (iG), for which exact samplers are available.

For illustration purposes, we choose a simulation setting akin to an example in [5], where one

[4] Given a draw for s, a subset S of size s is drawn uniformly. Set $\theta_j = 0$ for all $j \notin S$ and draw $\theta_j, j \in S$ i.i.d. from standard Laplace.

[5] The beta-bernoulli priors in (1) induce a similar prior on the subset size.

Table 2: Squared error comparison over 100 replicates. Average squared error for the posterior median reported for BL (Bayesian Lasso), HS (horseshoe) and DL (Dirichlet-Laplace) with $a = 1/n$ and $a = 1/2$ respectively.

n	**1000**					
A	**2**	**3**	**4**	**5**	**6**	**7**
BL	299.30	385.68	424.09	450.20	474.28	493.03
HS	306.94	353.79	270.90	205.43	182.99	168.83
$\text{DL}_{1/n}$	368.45	679.17	671.34	374.01	213.66	160.14
$\text{DL}_{1/2}$	267.83	315.70	266.80	213.23	192.98	177.20

has a single observation y from a $n = 1000$ dimensional $\text{N}_n(\theta_0, \text{I}_n)$ distribution, with $\theta_0[1:10] = 10, \theta_0[11:100] = A$, and $\theta_0[101:1000] = 0$. We then vary A from 2 to 7 and summarize the squared error averaged across 100 replicates in Table 2. We only compare the Bayesian shrinkage priors here; the squared error for the posterior median is tabulated. Table 2 clearly illustrates the need for prior elicitation in high dimensions according to the need, shrinking the noise vs. signal detection.

For visual illustration and comparison, we finally present the results from a single replicate in the first simulation setting with $n = 200$, $q_n = 10$ and $A = 7$ in Figure 1 & 2. The blue circles indicate the entries of y, while the red circles correspond to the posterior median of θ. The shaded region corresponds to a 95% point wise credible interval for θ.

6. PROOFS OF CONCENTRATION RESULTS IN SECTION 3

In this section, we develop non-asymptotic bounds to the prior concentration which are subsequently used to prove the posterior lower bound results. An important tool used throughout is a general version of Anderson's lemma [30], providing a concentration result for multivariate Gaussian distributions:

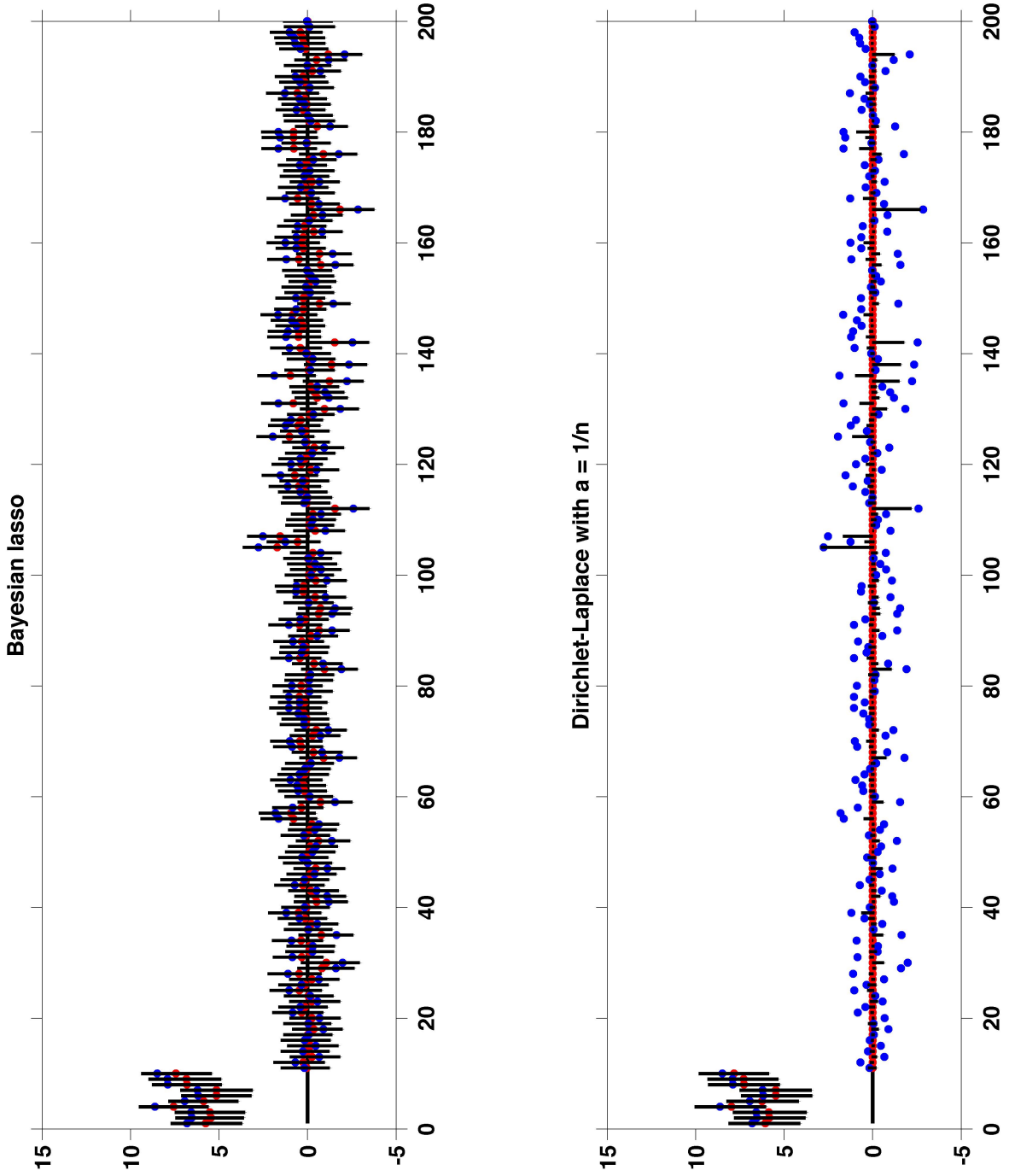

Figure 1: Simulation results from a single replicate with $n = 200, q_n = 10, A = 7$. Blue circles = entries of y, red circles = posterior median of θ, shaded region: 95% point wise credible interval for θ. Left panel: Bayesian lasso, right panel: $\text{DL}_{1/n}$ prior

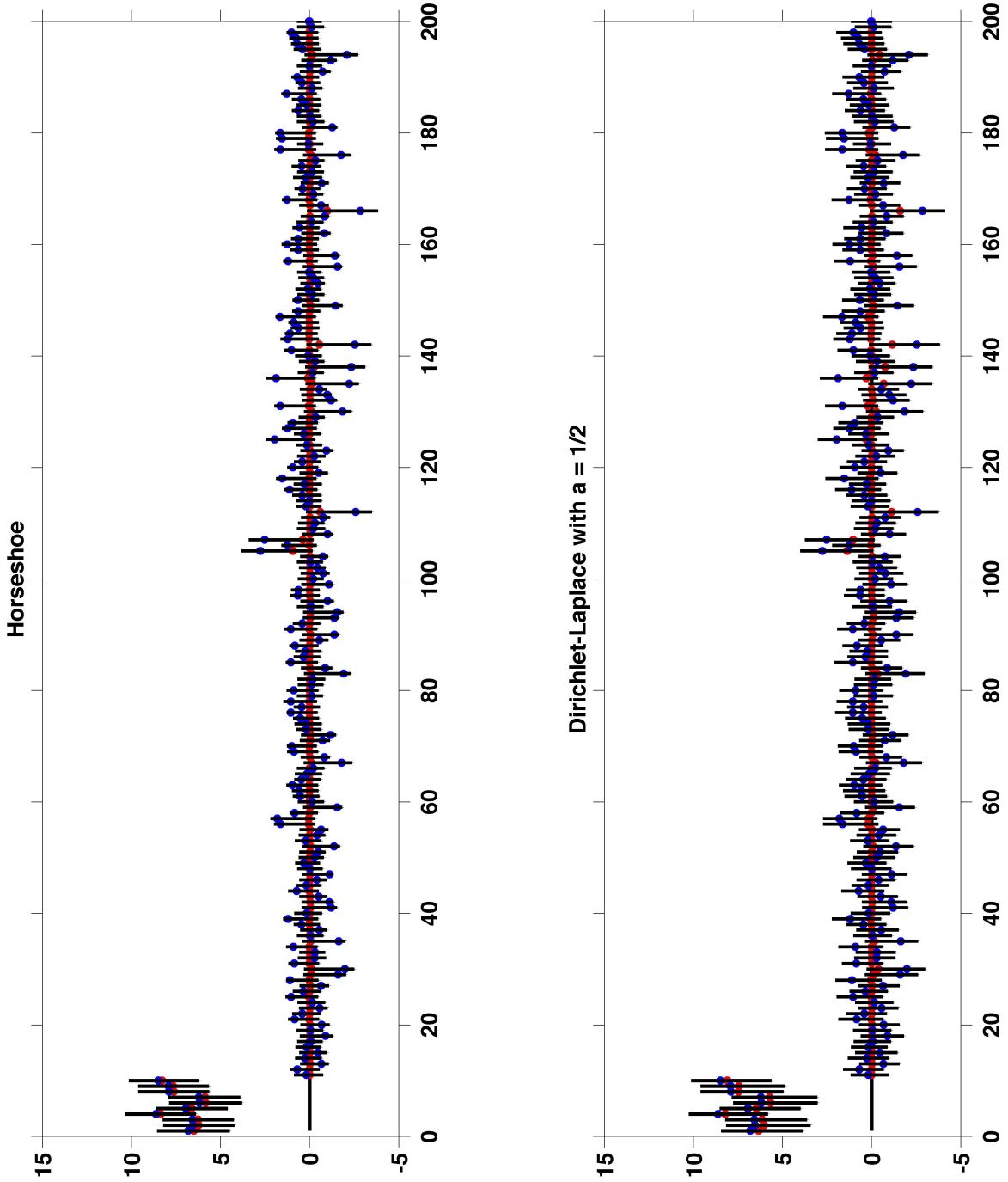

Figure 2: Simulation results from a single replicate with $n = 200, q_n = 10, A = 7$. Blue circles = entries of y, red circles = posterior median of θ, shaded region: 95% point wise credible interval for θ. Left panel: Horseshoe, right panel: $\text{DL}_{1/2}$ prior

Lemma 6.1. *Suppose $\theta \sim N_n(0, \Sigma)$ with Σ p.d. and $\theta_0 \in \mathbb{R}^n$. Let $\|\theta_0\|_{\mathbb{H}}^2 = \theta_0^{\mathrm{T}}\Sigma^{-1}\theta_0$. Then, for any $t > 0$,*

$$e^{-\frac{1}{2}\|\theta_0\|_{\mathbb{H}}^2}\mathbb{P}(\|\theta\|_2 \leq t/2) \leq \mathbb{P}(\|\theta - \theta_0\|_2 < t) \leq e^{-\frac{1}{2}\|\theta_0\|_{\mathbb{H}}^2}\mathbb{P}(\|\theta\|_2 < t).$$

It is well known that among balls of fixed radius, a zero mean multivariate normal distribution places the maximum mass on the ball centered at the origin. Lemma 6.1 provides a sharp bound on the probability of shifted balls in terms of the centered probability and the size of the shift, measured via the RKHS norm $\|\theta_0\|_{\mathbb{H}}^2$.

For GL shrinkage priors of the form (2), given $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)^{\mathrm{T}}$ and τ, the elements of θ are conditionally independent with $\theta \mid \boldsymbol{\psi}, \tau \sim \mathrm{N}_n(0, \Sigma)$ with $\Sigma = \mathrm{diag}(\psi_1\tau, \ldots, \psi_n\tau)$. Hence we can use Lemma 6.1 to obtain

$$\begin{aligned} &e^{-1/(2\tau)\sum_{j=1}^n \theta_{0j}^2/\psi_j}\,\mathbb{P}(\|\theta\|_2 < t_n/2 \mid \boldsymbol{\psi}, \tau) \leq \mathbb{P}(\|\theta - \theta_0\|_2 \leq t_n \mid \boldsymbol{\psi}, \tau) \\ &\leq e^{-1/(2\tau)\sum_{j=1}^n \theta_{0j}^2/\psi_j}\,\mathbb{P}(\|\theta\|_2 < t_n \mid \boldsymbol{\psi}, \tau). \end{aligned} \tag{19}$$

Letting $X_j = \theta_j^2$, X_j's are conditionally independent given $(\boldsymbol{\psi}, \tau)$ with X_j having a density $f(x_j \mid \boldsymbol{\psi}, \tau) = D/(\sqrt{\tau\psi_j x_j})e^{-x_j/(2\tau\psi_j)}$ on $(0, \infty)$, where $D = 1/(\sqrt{2\pi})$. Hence, with $w_n = t_n^2$,

$$\mathbb{P}(\|\theta\|_2 < t_n \mid \boldsymbol{\psi}, \tau) = D^n \int_{\sum x_j \leq w_n} \prod_{j=1}^n \frac{1}{\sqrt{x_j\tau\psi_j}} e^{-x_j/(2\tau\psi_j)} d\boldsymbol{x}. \tag{20}$$

For sake of brevity, we use $\{\sum x_j \leq w_n\}$ in (20) and all future references to denote the region $\{\boldsymbol{x} \in \mathbb{R}^n : x_j \geq 0\, \forall\, j = 1, \ldots, n,\, \sum_{j=1}^n x_j \leq w_n\}$. To estimate two-sided bounds for the marginal concentration $\mathbb{P}(\|\theta - \theta_0\|_2 \leq t_n)$, we need to combine (19) & (20) and integrate out $\boldsymbol{\psi}$ and τ carefully. We start by proving Theorem 3.2 & Theorem 3.3 where one only needs to integrate out τ.

6.1 **Proof of Theorem 3.2**

In (20), set $\psi_j = 1$ for all j, recall $D = 1/\sqrt{2\pi}$ and $w_n = t_n^2$, and integrate over τ to obtain,

$$\mathbb{P}(\|\theta\|_2 \le t_n) = D^n \int_{\tau=0}^{\infty} f(\tau) \Big[\int_{\sum x_j \le w_n} \prod_{j=1}^{n} \frac{1}{\sqrt{x_j \tau}} e^{-x_j/(2\tau)} d\boldsymbol{x} \Big] d\tau. \tag{21}$$

Substituting $f(\tau) = c\tau^{-(1+\alpha)} e^{-\beta/\tau}$ with $c = \beta^\alpha/\Gamma(\alpha)$ and using Fubini's theorem to interchange the order of integration between x and τ, (21) equals

$$\begin{aligned}
& cD^n \int_{\sum x_j \le w_n} \prod_{j=1}^{n} \frac{1}{\sqrt{x_j}} \Big[\int_{\tau=0}^{\infty} \tau^{-(1+n/2+\alpha)} e^{-\frac{1}{2\tau}(2\beta + \sum x_j)} d\tau \Big] d\boldsymbol{x} \\
&= cD^n 2^{n/2+\alpha} \Gamma(n/2+\alpha) \int_{\sum x_j \le w_n} \frac{1}{(2\beta + \sum x_j)^{n/2+\alpha}} \prod_{j=1}^{n} \frac{1}{\sqrt{x_j}} \, d\boldsymbol{x} \\
&= cD^n 2^{n/2+\alpha} w_n^{n/2} \Gamma(n/2+\alpha) \int_{\sum x_j \le 1} \frac{1}{(2\beta + w_n \sum x_j)^{n/2+\alpha}} \prod_{j=1}^{n} \frac{1}{\sqrt{x_j}} \, d\boldsymbol{x}.
\end{aligned} \tag{22}$$

We now state the Dirichlet integral formula (4.635 in [12]) to simplify a class of integrals as above over the simplex Δ^{n-1}:

Lemma 6.2. *Let $h(\cdot)$ be a Lebesgue integrable function and $\alpha_j > 0, j = 1, \ldots, n$. Then,*

$$\int_{\sum x_j \le 1} h\big(\sum x_j\big) \prod_{j=1}^{n} x_j^{\alpha_j - 1} d\boldsymbol{x} = \frac{\prod_{j=1}^{n} \Gamma(\alpha_j)}{\Gamma\big(\sum_{j=1}^{n} \alpha_j\big)} \int_{t=0}^{1} h(t)\, t^{(\sum \alpha_j) - 1} dt.$$

Lemma 6.2 follows simply by noting that the left hand side is $\mathbb{E} h(\sum_{j=1}^{n} X_j)$ up to normalizing constants where $(X_1, \ldots, X_n) \sim \text{Diri}(\alpha_1, \ldots, \alpha_n, 1)$, so that $\sum_{j=1}^{n} X_j \sim \text{Beta}(\sum \alpha_j, 1)$. Such probabilistic intuitions will be used later to reduce more complicated integrals over a simplex to a single integral on $(0, 1)$.

Lemma 6.2 with $h(t) = 1/(2\beta + w_n t)^{n/2+\alpha}$ applied to (22) implies

$$\mathbb{P}(\|\theta\|_2 \le t_n) = cD^n 2^{n/2+\alpha} w_n^{n/2} \Gamma(n/2+\alpha) \frac{\Gamma(1/2)^n}{\Gamma(n/2)} \int_{t=0}^{1} \frac{t^{n/2-1}}{(2\beta + w_n t)^{n/2+\alpha}} dt. \tag{23}$$

Substituting $D = 1/\sqrt{2\pi}$, bounding $(2\beta + w_n t)^{n/2+\alpha} \ge (2\beta)^{\alpha+1} (2\beta + w_n t)^{n/2-1}$, and letting

$\tilde{w}_n = w_n/(2\beta)$, (23) can be bounded above by

$$\frac{\Gamma(n/2+\alpha)}{\Gamma(n/2)\Gamma(\alpha)(2\beta)^{\alpha+1}}\tilde{w}_n^{n/2}\int_{t=0}^{1}\frac{t^{n/2-1}}{(1+\tilde{w}_n t)^{n/2-1}}dt \leq \frac{w_n\Gamma(n/2+\alpha)}{\Gamma(n/2)\Gamma(\alpha)(2\beta)^{\alpha+1}}\left(\frac{\tilde{w}_n}{1+\tilde{w}_n}\right)^{n/2-1},$$

where the second inequality above uses $t/(a+t)$ is an increasing function in $t > 0$ for fixed $a > 0$. By definition, $w_n = n^\delta$ for $0 < \delta < 1$ and hence $\frac{w_n\Gamma(n/2+\alpha)}{\Gamma(n/2)\Gamma(\alpha)(2\beta)^{\alpha+1}}$ can be bounded above by $e^{C_1 \log n}$. Also, using $(1-x)^{1/x} \leq e$ for all $x > 0$, $\{\tilde{w}_n/(1+\tilde{w}_n)\}^{n/2-1}$ can be bound above by $e^{-C_2 n/w_n} = e^{-C_2 n^{1-\delta}}$. Hence the overall bound is $e^{-Cn^{1-\delta}}$ for some appropriate constant $C > 0$. □

6.2 Proof of Theorem 3.3

We start with the upper bound in (7). The steps are similar as above and hence only a sketch is provided. Bounding $f(\tau) \leq M$ and interchanging order of integrals in (21),

$$\mathbb{P}(\|\theta\|_2 \leq t_n) \leq MD^n 2^{n/2-1}\Gamma(n/2-1)w_n \int_{\sum x_j \leq 1} \frac{1}{(\sum x_j)^{n/2-1}}\prod_{j=1}^{n}\frac{1}{\sqrt{x_j}}\,d\boldsymbol{x}. \tag{24}$$

Invoking Lemma 6.2 with $h(t) = (1/t)^{n/2-1}$ in (24), the upper bound in (7) is proved:

$$MD^n 2^{n/2-1}\Gamma(n/2-1)w_n\frac{\Gamma(1/2)^n}{\Gamma(n/2)}\int_{x=0}^{1} x^{n/2-1}/x^{n/2-1}dx = (M/2)\frac{w_n}{n/2-1} = C_2 n^{-(1-\delta)}.$$

We turn towards proving the lower bound to the centered concentration in (7). Recalling that $f(\tau) \geq 1/M$ on $(0,1)$ for $f \in \mathcal{F}$, and interchanging integrals in (21), we have, with $K = 1/M$,

$$\mathbb{P}(\|\theta\|_2 \leq t_n) \geq KD^n \int_{\sum x_j \leq w_n}\prod_{j=1}^{n}\frac{1}{\sqrt{x_j}}\left[\int_{\tau=0}^{1}\tau^{-n/2}e^{-\sum x_j/(2\tau)}d\tau\right]d\boldsymbol{x}. \tag{25}$$

We state Lemma 6.3 to lower bound the inner integral over τ; a proof can be found in the Appendix. Recall $\int_{\tau=0}^{\infty}\tau^{-n/2}e^{-a_n/(2\tau)}d\tau = \Gamma(n/2-1)(2/a_n)^{n/2-1}$. Lemma 6.3 shows that the same integral over $(0,1)$ is of the same order when $a_n \precsim n$.

Lemma 6.3. *For a sequence $a_n \leq n/(2e)$, $\int_{\tau=0}^{1}\tau^{-n/2}e^{-a_n/(2\tau)}d\tau \geq (2/a_n)^{n/2-1}\Gamma(n/2-1)\xi_n$,*

where $\xi_n \uparrow 1$ with $(1-\xi_n) \leq D/\sqrt{n}$ for some constant $D > 0$.

Clearly $\sum x_j \leq w_n$ and hence we can apply Lemma 6.3 in (25) to get

$$\mathbb{P}(\|\theta\|_2 \leq t_n) \geq K\xi_n D^n 2^{n/2-1}\Gamma(n/2-1)w_n \int_{\sum x_j \leq 1} \frac{1}{(\sum x_j)^{n/2-1}} \prod_{j=1}^{n} \frac{1}{\sqrt{x_j}}\, d\boldsymbol{x}. \tag{26}$$

The rest of the proof proceeds exactly as in the upper bound case from (24) onwards. □

Finally, we combine Anderson's inequality (19) with (20) (with $\psi_j = 1$ for all j in this case) to bound the non-centered probability in (8). For the upper bound, we additionally use $f(\tau) \leq M$ for all τ to obtain

$$\mathbb{P}(\|\theta - \theta_0\|_2 \leq t_n) \leq MD^n \int_{\sum x_j \leq w_n} \prod_{j=1}^{n} \frac{1}{\sqrt{x_j}} \Big[\int_{\tau=0}^{\infty} \tau^{-n/2} e^{-[\|\theta_0\|_2^2 + \sum x_j]/(2\tau)} d\tau \Big] d\boldsymbol{x} \tag{27}$$

$$= MD^n 2^{n/2-1}\Gamma(n/2-1)w_n^{n/2} \int_{\sum x_j \leq 1} \frac{1}{(\|\theta_0\|_2^2 + w_n \sum x_j)^{n/2-1}} \prod_{j=1}^{n} \frac{1}{\sqrt{x_j}}\, d\boldsymbol{x} \tag{28}$$

$$= MD^n 2^{n/2-1}\Gamma(n/2-1)w_n^{n/2} \frac{\Gamma(1/2)^n}{\Gamma(n/2)} \int_{x=0}^{1} \frac{x^{n/2-1}}{(\|\theta_0\|_2^2 + w_n x)^{n/2-1}} dx. \tag{29}$$

In the above display, (28) - (29) follows from applying Lemma 6.2 with $h(t) = 1/(\|\theta_0\|_2^2 + w_n t)^{n/2-1}$. Simplifying constants in (29) as before and using $t/(a+t)$ is an increasing function in $t > 0$ for fixed $a > 0$, we complete the proof by bounding (29) above by

$$\frac{Cw_n}{(n/2-1)} \int_{x=0}^{1} \frac{(w_n x)^{n/2-1}}{(\|\theta_0\|_2^2 + w_n x)^{n/2-1}} dx \leq \frac{Cw_n}{(n/2-1)} \left(\frac{w_n}{w_n + \|\theta_0\|_2^2} \right)^{n/2-1} \leq \frac{Cw_n}{(n/2-1)} \left(\frac{w_n}{\|\theta_0\|_2^2} \right)^{n/2-1}.$$

The right hand side of the above display can be bounded above by $e^{-cn \log a_n}$ for some constant $c > 0$. Remark (3.4) readily follows from the above display; we didn't use the condition on $\|\theta_0\|_2$ so far.

For the lower bound on the prior concentration in the non-centered case, we combine Anderson's inequality (19) in the reverse direction along with (20). We then use the same trick as in the centered case to restrict the integral over τ to $(0, 1)$ in (30). Note that the integral over the x's is

over $\sum x_j \leq v_n$ with $v_n = t_n^2/4$ as a consequence of (19). Hence,

$$\mathbb{P}(\|\theta - \theta_0\|_2 \leq t_n) \geq KD^n \int_{\sum x_j \leq v_n} \prod_{j=1}^{n} \frac{1}{\sqrt{x_j}} \Big[\int_{\tau=0}^{1} \tau^{-n/2} e^{-[\|\theta_0\|_2^2 + \sum x_j]/(2\tau)} d\tau \Big] d\boldsymbol{x}. \tag{30}$$

Noting that $\|\theta_0\|_2^2 + \sum x_j \leq \|\theta_0\|_2^2 + v_n = o(n)$, we can invoke Lemma 6.3 to lower bound the inner integral over τ by $\xi_n \Gamma(n/2-1) 2^{n/2-1}/(\|\theta_0\|_2^2 + \sum x_j)^{n/2-1}$ and proceed to obtain the same expressions as in (28) & (29) with M replaced by $K\xi_n$ and w_n by v_n. The proof is then completed by observing that the resulting lower bound can be further bounded below as follows:

$$\begin{aligned} &\frac{Cv_n}{(n/2-1)} \int_{x=0}^{1} \frac{(v_n x)^{n/2-1}}{(\|\theta_0\|_2^2 + v_n x)^{n/2-1}} dx \geq \frac{Cv_n}{(n/2-1)} \int_{x=1/2}^{1} \frac{(v_n x)^{n/2-1}}{(\|\theta_0\|_2^2 + v_n x)^{n/2-1}} dx \\ &\geq \frac{Cv_n}{(n/2-1)} \Big(\frac{v_n/2}{(\|\theta_0\|_2^2 + v_n/2)} \Big)^{n/2-1} \geq \frac{Cv_n}{(n/2-1)} \Big(\frac{v_n/2}{2\|\theta_0\|_2^2} \Big)^{n/2-1}, \end{aligned}$$

where the last inequality uses $t_n \leq \|\theta_0\|_2$ so that $\|\theta_0\|_2^2 + v_n \leq 2\|\theta_0\|_2^2$. □

We state and prove a result on concentration of GL priors in Theorem 6.4, with the proof of Theorem 3.5 following as a corollary of this more general result. The steps for obtaining Theorem 3.5 from Theorem 6.4 can be found in the proof of Theorem 3.7.

Theorem 6.4. *Assume $\theta \sim \mathrm{GL}$ with $f \in \mathcal{F}$ and $g \equiv \mathrm{Exp}(\lambda)$ for some constant $\lambda > 0$. Also assume θ_0 has only one non-zero entry. Let $w_n = t_n^2$. Then, for a global constant $C_1 > 0$ depending only on M in the definition of $\mathcal{F}$,*

$$\mathbb{P}(\|\theta - \theta_0\|_2 \leq t_n) \leq C_1 \int_{\psi_1=0}^{\infty} \frac{\psi_1^{(n-3)/2}}{\big\{\psi_1 + \|\theta_0\|_2^2/(\pi w_n)\big\}^{(n-3)/2}} e^{-\psi_1} d\psi_1. \tag{31}$$

Let $v_n = r_n^2/4$ satisfy $v_n = O(\sqrt{n})$. Then, for $\|\theta_0\|_2 \geq 1/\sqrt{n}$,

$$\mathbb{P}(\|\theta - \theta_0\|_2 \leq r_n) \geq C_2 e^{-d_2\sqrt{n}} \int_{\psi_1 = c_1\|\theta_0\|_2^2}^{\infty} \frac{\psi_1^{(n-3)/2}}{\big\{\psi_1 + \|\theta_0\|_2^2/(\pi v_n)\big\}^{(n-3)/2}} e^{-\psi_1} d\psi_1, \tag{32}$$

where c_1, d_2, C_2 are positive global constants with $c_1 \geq 2$ and C_2 depends only on M in the definition of $\mathcal{F}$.

6.3 **Proof of Theorem 6.4**

Without loss of generality, we assume g to be the $\text{Exp}(1)$ distribution since the rate parameter λ can be absorbed into the global parameter τ with the resulting distribution still in $\mathcal{F}$. Also, assume the only non-zero entry in θ_0 is θ_{01}, so that $\|\theta_0\|_2^2 = |\theta_{01}|^2$. The steps of the proof follow the same structure as in Theorem 3.3, i.e., using Anderson's inequality to bound the non-centered concentration given $\boldsymbol{\psi}, \tau$ by the centered concentration as in (19) and exploiting the properties of $\mathcal{F}$ to ensure that the bounds are tight. A substantial additional complication arises in integrating out $\boldsymbol{\psi}$ in this case, requiring evaluation of complicated multiple integrals (Lemma 6.5) and subsequent analysis.

We start with the upper bound (31). Combining (19) & (20), and bounding $f(\tau) \leq M$ yields:

$$
\begin{aligned}
&\mathbb{P}(\|\theta - \theta_0\|_2 \leq t_n) \\
&\leq D^n \int_{\tau=0}^{\infty} f(\tau) e^{-1/(2\tau)\sum_{j=1}^n \theta_{0j}^2/\psi_j} \int_{\boldsymbol{\psi}} g(\boldsymbol{\psi}) \Bigg[\int_{\sum x_j \leq w_n} \prod_{j=1}^n \frac{1}{\sqrt{x_j \tau \psi_j}} e^{-x_j/(2\tau\psi_j)} d\boldsymbol{x} \Bigg] d\boldsymbol{\psi} d\tau \\
&\leq M D^n \int_{\sum x_j \leq w_n} \prod_{j=1}^n \frac{1}{\sqrt{x_j}} \int_{\boldsymbol{\psi}} \prod_{j=1}^n \frac{g(\psi_j)}{\sqrt{\psi_j}} \Bigg[\int_{\tau=0}^{\infty} \tau^{-n/2} e^{-\frac{1}{2\tau}\left[\theta_{01}^2/\psi_1 + \sum x_j/\psi_j\right]} d\tau \Bigg] d\boldsymbol{\psi} d\boldsymbol{x} \\
&= M D^n 2^{n/2-1} \Gamma(n/2-1) w_n^{n/2} \int_{\boldsymbol{\psi}} \prod_{j=1}^n \frac{g(\psi_j)}{\sqrt{\psi_j}} \Bigg[\int_{\sum x_j \leq 1} \frac{\prod_{j=1}^n x_j^{-1/2}}{[\|\theta_0\|_2^2/\psi_1 + w_n \sum x_j/\psi_j]^{n/2-1}} \, d\boldsymbol{x} \Bigg] d\boldsymbol{\psi}.
\end{aligned} \tag{33}
$$

Compare (33) with (28). The crucial difference in this case is that the inner integral over the simplex $\sum_{j=1}^n x_j \leq 1$ is no longer a function of $\sum_{j=1}^n x_j$, rendering Lemma 6.2 inapplicable. An important technical contribution of this paper in Lemma 6.5 below is that complicated multiple integrals over the simplex as above can be reduced to a single integral over $(0, 1)$:

Lemma 6.5. *Let $q_j, j = 0, 1, \ldots, n$ be positive numbers. Then,*

$$
\int_{\sum x_j \leq 1} \frac{\prod_{j=1}^n x_j^{-1/2}}{[\sum_{j=1}^n q_j x_j + q_0]^{n/2-1}} d\boldsymbol{x} = \frac{\Gamma(1/2)^n}{\Gamma(n/2)} q_0 (n/2-1) \int_{x=0}^1 \frac{x^{n/2-2}(1-x)}{\prod_{j=1}^n \sqrt{q_j x + q_0}} dx.
$$

A proof of Lemma 6.5 can be found in the Appendix, which utilizes a beautiful identity found

in [7]. We didn't find any reference for Lemma 6.5, though a related integral with $n/2$ in the exponent in the denominator appears in [12].

Applying Lemma 6.5 with $q_0 = \|\theta_0\|_2^2/\psi_1$ and $q_j = w_n/\psi_j$ to evaluate the inner integral over x, (33) equals

$$(M\|\theta_0\|_2^2/2)w_n^{n/2}\int_{\psi}\left[\prod_{j=1}^{n}\frac{g(\psi_j)}{\sqrt{\psi_j}}\right]\frac{1}{\psi_1}\int_{x=0}^{1}\frac{x^{n/2-2}(1-x)}{\prod_{j=1}^{n}\sqrt{(w_nx/\psi_j+q_0)}}dxd\boldsymbol{\psi}, \tag{34}$$

noting that $(n/2-1)D^n2^{n/2-1}\Gamma(n/2-1)\,\Gamma(1/2)^n/\Gamma(n/2) = 1/2$.

So, at this point, we are down from the initial $(2n+1)$ integrals to $(n+1)$ integrals. Next, using $g(\psi_j) = e^{-\psi_j}1(\psi_j > 0)$ to integrate out $\psi_j, j = 2,\ldots,n$, (34) equals

$$(M\|\theta_0\|_2^2/2)w_n^{n/2}\int_{\psi_1=0}^{\infty}\frac{e^{-\psi_1}}{\psi_1\sqrt{\psi_1}}\int_{x=0}^{1}\frac{x^{n/2-2}(1-x)}{\sqrt{w_nx/\psi_1+q_0}}\left\{\int_{\psi=0}^{\infty}\frac{e^{-\psi}}{\sqrt{w_nx+\psi q_0}}d\psi\right\}^{n-1}dx\,d\psi_1. \tag{35}$$

Using a standard identity and Lemma 6.6 below (proof in Appendix),

$$\int_{\psi=0}^{\infty}\frac{e^{-\psi}}{\sqrt{w_nx+\psi q_0}}d\psi = \frac{\sqrt{\pi}}{\sqrt{q_0}}\exp(w_nx/q_0)\text{erfc}(\sqrt{w_nx/q_0}) \le \frac{1}{\sqrt{w_nx+q_0/\pi}}.$$

Lemma 6.6. *Let erfc$(x) = \frac{2}{\sqrt{\pi}}\int_x^{\infty}e^{-t^2}dt$ denote the complementary error function. Then,*

$$\sqrt{\pi}e^x\text{erfc}(\sqrt{x}) \le \frac{1}{\sqrt{x+1/\pi}} \tag{36}$$

$$\sqrt{\pi}e^x\text{erfc}(\sqrt{x}) \ge \left\{\frac{1}{\sqrt{x}}\right\}^{1+\delta} \tag{37}$$

where (37) *holds for any $\delta > 0$ provided $x \ge 2$.*

Hence, the expression in (35) can be bounded above by

$$(M/2)\left\|\theta_0\right\|_2^2 w_n^{n/2}\int_{\psi_1=0}^{\infty}\frac{e^{-\psi_1}}{\psi_1}\int_{x=0}^{1}\frac{x^{n/2-2}(1-x)}{\sqrt{\left(w_n x+\left\|\theta_0\right\|_2^2\right)}\left[w_n x+\left\|\theta_0\right\|_2^2/(\pi\psi_1)\right]^{(n-1)/2}}dx\,d\psi_1$$
$$=(M/2)\left\|\theta_0\right\|_2^2 w_n^{n/2}\int_{\psi_1=0}^{\infty}e^{-\psi_1}\psi_1^{(n-3)/2}\int_{x=0}^{1}\frac{x^{n/2-2}(1-x)}{\sqrt{\left(w_n x+\left\|\theta_0\right\|_2^2\right)}\left[w_n x\psi_1+\left\|\theta_0\right\|_2^2/\pi\right]^{(n-1)/2}}dx\,d\psi_1. \tag{38}$$

Let us aim to bound the inner integral over x in (38). We upper bound $(1-x)$ in the numerator by 1, lower-bound $\sqrt{(w_n x+\left\|\theta_0\right\|_2^2)}$ in the denominator by $\sqrt{\left\|\theta_0\right\|_2^2}$ and multiply a $\sqrt{w_n x\psi_1+\left\|\theta_0\right\|_2^2/\pi}$ term in the numerator and denominator to get

$$\int_{x=0}^{1}\frac{x^{n/2-2}(1-x)}{\sqrt{w_n x+\left\|\theta_0\right\|_2^2}\left[w_n x\psi_1+\left\|\theta_0\right\|_2^2/\pi\right]^{(n-1)/2}}dx$$
$$\leq\frac{\sqrt{w_n\psi_1+\left\|\theta_0\right\|_2^2/\pi}}{\sqrt{\left\|\theta_0\right\|_2^2}}\int_{x=0}^{1}\frac{x^{n/2-2}}{\left(w_n x\psi_1+\left\|\theta_0\right\|_2^2/\pi\right)^{n/2}}dx.$$

We use the fact that $\int_{x=0}^{1}x^{n/2-2}/(\alpha x+\beta)^{n/2}dx=2(\alpha+\beta)^{1-n/2}/\{\beta(n-2)\}$ to conclude that the last line in the above display equals

$$\frac{\sqrt{w_n\psi_1+\left\|\theta_0\right\|_2^2/\pi}}{\sqrt{\left\|\theta_0\right\|_2^2}}\frac{2\pi}{\left\|\theta_0\right\|_2^2}\frac{\left(w_n\psi_1+\left\|\theta_0\right\|_2^2/\pi\right)^{1-n/2}}{(n-2)}$$
$$=\frac{1}{\sqrt{\left\|\theta_0\right\|_2^2}}\frac{\pi}{\left\|\theta_0\right\|_2^2(n/2-1)}\left(w_n\psi_1+\left\|\theta_0\right\|_2^2/\pi\right)^{-(n-3)/2}.$$

Substituting this in (38), we finally obtain:

$$\mathbb{P}(\left\|\theta-\theta_0\right\|_2\leq t_n)\leq\frac{C_1 w_n}{(n/2-1)}\sqrt{\frac{w_n}{\left\|\theta_0\right\|_2^2}}\int_{\psi_1=0}^{\infty}\frac{\psi_1^{(n-3)/2}}{\left\{\psi_1+\left\|\theta_0\right\|_2^2/(\pi w_n)\right\}^{(n-3)/2}}e^{-\psi_1}d\psi_1, \tag{39}$$

where $C_1>0$ is a global constant (depending only on M). (31) clearly follows from (39). □

Lower bound: We proceed to obtain a lower bound to $\mathbb{P}(\left\|\theta-\theta_0\right\|_2<r_n)$ similar to (39) under

additional assumptions on r_n as in the statement of Theorem 6.4. To that end, note that in the proof of the upper bound here, we used only two inequalities until (34): (i) Anderson's inequality in (19) and (ii) upper bounding $f(\tau)$ by M. As in the proof of the lower bound in Theorem 3.3, we obtain a lower bound similar to the expression in (34) by (i) using Anderson's inequality (19) in the reverse direction, and (ii) using $f(\tau) \geq K = 1/M$ on $(0,1)$:

$$\begin{aligned}&\mathbb{P}(\|\theta - \theta_0\|_2 \leq r_n) \\ &\geq KD^n \int_{\sum x_j \leq v_n} \prod_{j=1}^n \frac{1}{\sqrt{x_j}} \int_\psi \prod_{j=1}^n \frac{g(\psi_j)}{\sqrt{\psi_j}} \left[\int_{\tau=0}^1 \tau^{-n/2} e^{-\frac{1}{2\tau}\left[\|\theta_0\|_2^2/\psi_1 + \sum_{j=1}^n x_j/\psi_j\right)} d\tau \right] d\boldsymbol{\psi}\, d\boldsymbol{x}. \end{aligned} \tag{40}$$

However, unlike Theorem 3.3, we cannot directly resort to Lemma 6.3 since $a_n = \|\theta_0\|_2^2/\psi_1 + \sum_{j=1}^n x_j/\psi_j$ can be arbitrarily large if ψ_j's are close enough to zero. This necessitates a more careful analysis in bounding below the expression in (40) by constraining the ψ_j's to an appropriate region Γ away from zero:

$$\Gamma = \left\{ c_1 \|\theta_0\|_2^2 \leq \psi_1 \leq c_2 \|\theta_0\|_2^2,\, \psi_j \geq c_3/\sqrt{n},\, j = 2, \ldots, n \right\}.$$

In the above display, $c_1 < c_2$ and $c_3 > 1$ are positive constants to be chosen later, that satisfy

$$1/c_1 + \max\{1/(c_1 \|\theta_0\|_2^2), \sqrt{n}/c_3\} v_n \leq n/(2e). \tag{41}$$

With (41), we can invoke Lemma 6.3 to bound below the integral over τ in (40), since for $\psi \in \Gamma$,

$$\begin{aligned}&\|\theta_0\|_2^2/\psi_1 + \sum_{j=1}^n x_j/\psi_j \leq 1/c_1 + \max\{1/(c_1 \|\theta_0\|_2^2), \sqrt{n}/c_3\} \sum_{j=1}^n x_j \\ &\leq 1/c_1 + \max\{1/(c_1 \|\theta_0\|_2^2), \sqrt{n}/c_3\} v_n \leq n/(2e)\end{aligned}$$

by (41). The resulting lower bound is exactly same as (33) with M replaced by $K\xi_n$ and w_n by v_n, where $\xi_n \uparrow 1$ is as in Lemma 6.3. As in the upper bound calculations (33) - (34), we invoke Lemma 6.5 with $q_0 = \|\theta_0\|_2^2/\psi_1$ and $q_j = v_n/\psi_j$ to reduce the multiple integral over the simplex

and bound the expression in (40) below by

$$(K\left\|\theta_0\right\|_2^2/2)\xi_n v_n^{n/2}\int_{\psi\in\Gamma}\left[\prod_{j=1}^{n}\frac{g(\psi_j)}{\sqrt{\psi_j}}\right]\frac{1}{\psi_1}\int_{x=1/2}^{3/4}\frac{x^{n/2-2}(1-x)}{\prod_{j=1}^{n}\sqrt{(v_n x/\psi_j+q_0)}}dxd\psi =$$
$$(K\left\|\theta_0\right\|_2^2/2)\xi_n v_n^{n/2}\int_{\psi_1=c_1\|\theta_0\|^2}^{c_2\|\theta_0\|^2}\frac{e^{-\psi_1}}{\psi_1}\int_{x=1/2}^{3/4}\frac{x^{n/2-2}(1-x)}{\sqrt{v_n x+q_0\psi_1}}\left\{\int_{\psi=c_3/\sqrt{n}}^{\infty}\frac{e^{-\psi}}{\sqrt{v_n x+\psi q_0}}d\psi\right\}^{n-1}dx\,d\psi_1. \tag{42}$$

Note the inner integral over x is restricted to $(1/2, 3/4)$. Now,

$$\int_{\psi=c_3/\sqrt{n}}^{\infty}\frac{e^{-\psi}}{\sqrt{v_n x+\psi q_0}}d\psi=\frac{\sqrt{\pi}}{\sqrt{q_0}}e^{v_n x/q_0}\text{erfc}\left(\sqrt{v_n x/q_0+c_3/\sqrt{n}}\right). \tag{43}$$

Since we have restricted $x\geq 1/2$ in (42) and $v_n\psi_1/\left\|\theta_0\right\|_2^2\geq c_1 v_n$ on Γ, we have $\sqrt{v_n x/q_0+c_3/\sqrt{n}}\geq\sqrt{c_1}$ provided $v_n\geq 1$. Thus, choosing $c_1>2$, we use the lower bound to the erfc function from Lemma 6.6 to bound the expression in the r.h.s. of (43) as:

$$\frac{\sqrt{\pi}}{\sqrt{q_0}}e^{v_n x/q_0}\text{erfc}\left(\sqrt{v_n x/q_0+c_3/\sqrt{n}}\right)\geq\frac{1}{\sqrt{q_0}}e^{-c_3/\sqrt{n}}\left[\frac{1}{\sqrt{v_n x/q_0+c_3/\sqrt{n}}}\right]^{1+\delta}$$
$$\geq\frac{1}{\sqrt{q_0}}e^{-c_3/\sqrt{n}}\frac{1}{\sqrt{v_n x/q_0+3/(4\pi)}}\frac{1}{(1+c_2)^{\delta}}=\frac{e^{-c_3/\sqrt{n}}}{(1+c_2)^{\delta}}\frac{1}{\sqrt{v_n x+3q_0/(4\pi)}}.$$

In the second to third step, we used that $v_n x/q_0+c_3/\sqrt{n}\leq v_n x/q_0+3/(4\pi)$ for n larger than some constant. We choose $\delta=1/(n-1)$ and substitute the above lower bound for the l.h.s. of (43) into (42). This allows us to bound (42) below by

$$C_1\xi_n\left\|\theta_0\right\|_2^2 e^{-c_3(n-1)/\sqrt{n}}v_n^{n/2}\times$$
$$\int_{\psi_1=c_1\|\theta_0\|^2}^{c_2\|\theta_0\|^2}e^{-\psi_1}\psi_1^{(n-3)/2}\int_{x=1/2}^{3/4}\frac{x^{n/2-2}(1-x)}{\sqrt{\left(v_n x+\left\|\theta_0\right\|_2^2\right)}\left[v_n\psi_1 x+3\left\|\theta_0\right\|_2^2/(4\pi)\right]^{(n-1)/2}}dx\,d\psi_1. \tag{44}$$

Let us tackle the integral over x in (44). To that end, we first lower-bound $(1-x)$ in the numerator

by $1/4$, upper-bound $\sqrt{v_n x + \|\theta_0\|_2^2}$ in the denominator by $\sqrt{v_n + \|\theta_0\|_2^2}$. Next, we use the formula

$$\int_{x=1/2}^{3/4} \frac{x^{n/2-2}}{(\alpha x+\beta)^{n/2}} dx = \frac{2(\alpha+4\beta/3)^{1-n/2}}{\beta(n-2)} \left[1 - \left\{\frac{\alpha+4\beta/3}{\alpha+2\beta}\right\}^{n/2-1}\right],$$

with $\alpha = v_n\psi_1$ and $\beta = 3\|\theta_0\|_2^2/(4\pi)$. Now, $(\alpha+4\beta/3)/(\alpha+2\beta) = 1 - 2\beta/\{3(\alpha+2\beta)\}$ is bounded away from 0 and 1 since $c_1\|\theta_0\|_2^2 \le \alpha \le c_2\|\theta_0\|_2^2$. Thus,

$$\left[1 - \left\{\frac{\alpha+4\beta/3}{\alpha+2\beta}\right\}^{n/2-1}\right] \ge 1/2$$

for n large. Substituting all these in (44), we finally obtain:

$$\mathbb{P}(\|\theta-\theta_0\|_2 \le r_n) \ge \frac{C_2\xi_n v_n \exp(-c_3\sqrt{n})}{(n/2-1)} \sqrt{\frac{v_n}{v_n+\|\theta_0\|_2^2}} \int_{\psi_1=c_1\|\theta_0\|^2}^{c_2\|\theta_0\|^2} \frac{\psi_1^{(n-3)/2}}{\left\{\psi_1 + \|\theta_0\|_2^2/(\pi v_n)\right\}^{(n-3)/2}} e^{-\psi_1} d\psi_1, \tag{45}$$

where $C_2 > 0$ is a global constant depending only on K in the definition of $\mathcal{F}$ and $\xi_n \uparrow 1$ with $1-\xi_n \le D/\sqrt{n}$ for some constant $D > 0$. We only required $c_1 > 2$ so far. Since $\|\theta_0\|_2 \ge 1/\sqrt{n}$, choosing c_1 and c_3 to be sufficiently large constants, (41) can always be satisfied. The proof of (32) clearly follows from (45), since $\xi_n v_n/(n/2-1)\sqrt{\frac{v_n}{v_n+\|\theta_0\|_2^2}}$ can be bounded below by $e^{-c_4\sqrt{n}}$. □

6.4 **Proof of Theorem 3.7**

Let $m_n = (n-3)/2$. We set $t_n = s_n$, where s_n is the minimax rate corresponding to $q_n = 1$, so that $w_n = s_n^2 = \log n$. Also, let $\|\theta_0\|_2^2 = \pi w_n u_n^2$, where u_n is a slowly increasing sequence; we set $u_n = \log(\log n)$ for future references. Finally let $v_n = r_n^2/4 = \sqrt{m_n}$. With these choices, we proceed to show that (11) holds.

We first simplify (39) further. The function $x \to x/x(x+a)$ monotonically increases from 0

to 1 for any $a > 0$. Thus, for any $T_n > 0$,

$$\int_{\psi_1=0}^{\infty} \frac{\psi_1^{m_n}}{\left\{\psi_1 + \|\theta_0\|_2^2/(\pi w_n)\right\}^{m_n}} e^{-\psi_1} d\psi_1$$
$$\leq \int_{\psi_1=0}^{T_n} \frac{\psi_1^{m_n}}{\left\{\psi_1 + \|\theta_0\|_2^2/(\pi w_n)\right\}^{m_n}} e^{-\psi_1} d\psi_1 + \int_{\psi_1=T_n}^{\infty} e^{-\psi_1} d\psi_1 \leq \left(\frac{T_n}{T_n + u_n^2}\right)^{m_n} + e^{-T_n}. \quad (46)$$

We choose an appropriate T_n which gives us the necessary bound, namely $T_n = u_n\sqrt{m_n}$. Then, using the fact that $(1-x)^{1/x} \leq e^{-1}$ for all $x \in (0,1)$, we have

$$\left(\frac{T_n}{T_n + u_n^2}\right)^{m_n} = \left(\frac{\sqrt{m_n}}{\sqrt{m_n} + u_n}\right)^{m_n} = \left(1 - \frac{u_n}{\sqrt{m_n} + u_n}\right)^{m_n} \leq e^{-m_n u_n/(\sqrt{m_n}+u_n)} \leq e^{-u_n\sqrt{m_n}/2},$$

where for the last part used that $e^{-1/x}$ is an increasing function and $\sqrt{m_n} + u_n \leq 2\sqrt{m_n}$. Thus, substituting T_n in (46) yields, for a global constant $C_1 > 0$,

$$\mathbb{P}(\|\theta - \theta_0\|_2 \leq s_n) \leq \frac{C_1 w_n}{(n/2-1)} \sqrt{\frac{w_n}{\|\theta_0\|_2^2}}\, e^{-u_n\sqrt{m_n}/2}. \quad (47)$$

Next, again using the fact that $x \to x/x(x+a)$ is monotonically increasing, and choosing $c_2 = \infty$, we simplify the lower bound (45). Observe

$$\int_{\psi_1=c_1\|\theta_0\|^2}^{\infty} \frac{\psi_1^{m_n}}{\left\{\psi_1 + \|\theta_0\|_2^2/(\pi v_n)\right\}^{m_n}} e^{-\psi_1} d\psi_1$$
$$\geq \left(\frac{v_n}{v_n + C}\right)^{m_n} e^{-c_1\|\theta_0\|_2^2},$$

for some constant $C > 0$. Finally, using $(1-x)^{1/x} \geq e^{-2}$ for all $x \in (0, 1/2)$ and $e^{-1/x}$ is an increasing function in $x > 0$, we have,

$$\left(\frac{v_n}{v_n + C}\right)^{m_n} \geq e^{-\sqrt{m_n}/2}.$$

Hence, the integral is bounded below by $e^{-(\sqrt{m_n}+c_1\|\theta_0\|_2^2)/2}$, resulting in

$$\mathbb{P}(\|\theta-\theta_0\|_2 \le r_n) \ge \frac{C_2\xi_n v_n}{(n/2-1)}\sqrt{\frac{v_n}{v_n+\|\theta_0\|_2^2}}\; e^{-(\sqrt{m_n}+c_1\|\theta_0\|_2^2)/2}. \tag{48}$$

Thus, finally, noting that $u_n \to \infty$, (11) follows since

$$\frac{\mathbb{P}(||\theta-\theta_0||_2 < s_n)}{\mathbb{P}(||\theta-\theta_0||_2 < r_n)} \times e^{r_n^2} \le D\frac{w_n^{3/2}}{v_n} e^{C(\sqrt{m_n}+\sqrt{n}+\|\theta_0\|_2^2)}\, e^{-u_n\sqrt{m_n}/2} \to 0,$$

where $C, D > 0$ are constants. □

6.5 **Proof of Theorem 3.8**

As before, we assume $\lambda = 1$ w.l.g., since it can be absorbed in the constant appearing the sequence τ_n otherwise. As in the proof of Theorem 3.7, combine (19) & (20) to obtain

$$\begin{aligned}
\mathbb{P}(\|\theta-\theta_0\| < t_n) &\le D^n\tau_n^{-n/2}\int_\psi g(\psi)\Big\{\int_{\sum x_j \le w_n}\prod_{j=1}^n \frac{1}{\sqrt{x_j\psi_j}}\exp\Big(-\frac{x_j+\theta_{0j}^2}{2\psi_j}\Big)dx\Big\}d\psi \\
&= D^n\tau_n^{-n/2}w_n^{n/2}\int_{\sum x_j\le 1}\prod_{j=1}^n\frac{1}{\sqrt{x_j}}\Big\{\prod_{j=1}^n\int_{\psi_j=0}^\infty \frac{e^{-\psi_j}}{\sqrt{\psi_j}}\exp\Big(-\frac{w_n x_j+\theta_{0j}^2}{2\tau_n\psi_j}\Big)d\psi_j\Big\}dx,
\end{aligned}$$

where $w_n = t_n^2$. Using the fact $\int_0^\infty \frac{1}{\sqrt{x}}\exp\big\{-\big(\frac{a}{x}+x\big)\big\}dx = \sqrt{\pi}e^{-2\sqrt{a}}$, we obtain

$$\begin{aligned}
&\mathbb{P}(\|\theta-\theta_0\| < t_n) \\
&\le D^n\pi^{n/2}\tau_n^{-n/2}w_n^{n/2}\int_{\sum x_j\le 1}\prod_{j=1}^n\frac{1}{\sqrt{x_j}}\exp\Big\{-2\sqrt{\frac{w_n x_j+\theta_{0j}^2}{2\tau_n}}\Big\}dx \\
&\le \Big(\frac{D^2\pi w_n}{\tau_n}\Big)^{n/2}e^{-\frac{\sqrt{2}\|\theta_0\|_1}{\sqrt{\tau_n}}}\int_{\sum x_j\le 1}\prod_{j=1}^n\frac{1}{\sqrt{x_j}}dx = \Big(\frac{D^2\pi w_n}{\tau_n}\Big)^{n/2}e^{-\frac{\sqrt{2}\|\theta_0\|_1}{\sqrt{\tau_n}}}\frac{\Gamma(1/2)^n}{\Gamma(n/2+1)},
\end{aligned} \tag{49}$$

where the second to third inequality uses $x_j \geq 0$ and the last integral follows from Lemma 6.2. Along the same lines,

$$\begin{aligned}\mathbb{P}(\|\theta - \theta_0\| < r_n) &\geq \left(\frac{D^2\pi v_n}{\tau_n}\right)^{n/2} \int_{\sum x_j \leq 1} \prod_{j=1}^{n} \frac{1}{\sqrt{x_j}} \exp\left\{ -2\sqrt{\frac{v_n x_j + \theta_{0j}^2}{2\tau_n}} \right\} dx \\ &\geq \left(\frac{D^2\pi v_n}{\tau_n}\right)^{n/2} e^{-\frac{\sqrt{2}\|\theta_0\|_1 + \sqrt{nv_n}}{\sqrt{\tau_n}}} \frac{\Gamma(1/2)^n}{\Gamma(n/2+1)}, \end{aligned} \tag{50}$$

where $v_n = r_n^2/4$. From the second to third equation in the above display, we used $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $\sum_{j=1}^{n} \sqrt{x_j} \leq \sqrt{n}$ by Cauchy-Schwartz inequality if $x \in \Delta^{n-1}$. Thus, from (49) & (50), the ratio in (11) can be bounded above as:

$$\frac{\mathbb{P}(\|\theta - \theta_0\| < t_n)}{\mathbb{P}(\|\theta - \theta_0\| < r_n)} \leq \left(\frac{w_n}{v_n}\right)^{n/2} e^{\sqrt{2v_n n/\tau_n}}.$$

Choose $t_n = s_n$, $r_n = 2\sqrt{2}s_n$ so that $v_n = 2w_n = 2q_n \log(n/q_n)$ and $(w_n/v_n)^{n/2} = e^{-Cn}$. Clearly $v_n n/\tau_n \leq Cnq_n(\log n)^2$ and hence, $e^{\sqrt{2v_n n/\tau_n}} = o(e^{Cn})$ by assumption. Thus, the right hand side of the above display $\to 0$, proving the assertion of the Theorem.

Proof of Theorem 4.1

First, we will state a more general result on the concentration of $\text{DL}_{1/n}(\tau)$ when $\tau \sim \text{Exp}(\lambda)$. The result follows from a straightforward modification of Lemma 4.1 in [22] and the detailed proof is omitted here. Assume $\delta = t_n/(2n)$. For fixed numbers $0 < a < b < 1$, let

$$\eta_n = 1 - \frac{(n - q_n)\delta}{2q_n \log(n/q_n)} - \frac{(q_n - 1)b}{2q_n}, \quad \xi_n = 1 - \frac{(q_n - 1)a}{4q_n}.$$

Also, without loss of generality assume that $\{1\} \subset S_0 = \text{supp}(\theta_0)$, i.e., $\theta_{01} \neq 0$. Let $S_1 = S_0 \backslash \{1\}$. If $\theta_0 \in l_0(q_n; n)$, it follows that

$$P(\|\theta - \theta_0\| < t_n) \geq C\, \mathbb{P}(\tau \in [2q_n, 4q_n])\, A_n B_n,$$

where C is an absolute constant, and

$$\begin{aligned} A_n &= \exp\left\{ -q_n \log 2 - \sum_{j \in S_1} \frac{|\theta_{0j}|}{a} - \theta_{01}/(1-b)\eta_n \right\} \\ B_n &= \left[1 - \exp\left\{ -\frac{t_n}{\sqrt{2q_n}b} \right\}\right]^{q_n - 1} \left[1 - \exp\left\{ -\frac{t_n}{\sqrt{2q_n}(1-a/8)\xi_n} \right\}\right]. \end{aligned}$$

In our case, $|S_0| = 1$, $\theta_{01} = \sqrt{\log n}$, and $t_n = n^{\delta/2}$. Hence A_n is a constant, $B_n = \exp\{-K_1\sqrt{\log n}\}$ for some constant K_1 and $P(\tau \in [2q_n, 4q_n])$ is also a constant. Hence, under the assumptions of Theorem 4.1, $P(\|\theta - \theta_0\| < t_n) \geq \exp\{-C\sqrt{\log n}\}$.

7. ACKNOWLEDGEMENT

We thank Ismael Castillo and James Scott for sharing source code.

APPENDIX

Proof of Proposition 4.2

When $a = 1/n$, $\phi_j \sim \text{Beta}(1/n, 1 - 1/n)$ marginally. Hence, the marginal distribution of θ_j given τ is proportional to

$$\int_{\phi_j=0}^{1} e^{-|\theta_j|/(\phi_j \tau)} \left(\frac{\phi_j}{1-\phi_j}\right)^{1/n} \phi_j^{-2} d\phi_j.$$

Substituting $z = \phi_j/(1-\phi_j)$ so that $\phi_j = z/(1+z)$, the above integral reduces to

$$e^{-|\theta_j|/\tau} \int_{z=0}^{\infty} e^{-|\theta_j|/(\tau z)} z^{-(2-1/n)} dz \propto e^{-|\theta_j|/\tau} |\theta_j|^{1/n-1}.$$

In the general case, $\phi_j \sim \text{Beta}(a, (n-1)a)$ marginally. Substituting $z = \phi_j/(1-\phi_j)$ as before, the marginal density of θ_j is proportional to

$$e^{-|\theta_j|/\tau} \int_{z=0}^{\infty} e^{-|\theta_j|/(\tau z)} z^{-(2-a)} \left(\frac{1}{1+z}\right)^{na-1} dz.$$

The above integral can clearly be bounded below by a constant multiple of

$$e^{-|\theta_j|/\tau} \int_{z=0}^{1} e^{-|\theta_j|/(\tau z)} z^{-(2-a)} dz.$$

Resort to Lemma 6.3 to finish the proof.

Proof of Lemma 6.3

Using a simple change of variable,

$$\int_{\tau=0}^{1} \tau^{-n/2} e^{-a_n/(2\tau)} d\tau = \int_{z=1}^{\infty} z^{n/2-2} e^{-a_n z/2} dz$$
$$\left(\frac{2}{a_n}\right)^{n/2-1} \int_{t=a_n/2}^{\infty} t^{a_n/2-2} e^{-t} dt = \left(\frac{2}{a_n}\right)^{n/2-1} \left[\Gamma(n/2-1) - \int_{t=0}^{a_n/2} t^{n/2-2} e^{-t} dt\right]$$

Noting that $\int_{t=0}^{a_n/2} t^{n/2-2} e^{-t} dt \leq a_n^{n/2-1}/(n/2-1)$ and $a_n \leq n/(2e)$ by assumption, the last entry in the above display can be bounded below by

$$\left(\frac{2}{a_n}\right)^{n/2-1} \Gamma(n/2-1)\left[1 - \frac{a_n^{n/2-1}}{\Gamma(n/2)}\right] \geq \left(\frac{2}{a_n}\right)^{n/2-1} \left[1 - \frac{\{n/(2e)\}^{n/2-1}}{\Gamma(n/2)}\right].$$

Let $\xi_n = 1 - \{n/(2e)\}^{n/2-1}/\Gamma(n/2)$. Using the fact that $\Gamma(m) \geq \sqrt{2\pi}\, m^{m-1/2} e^{-m}$ for any $m > 0$, one has $\Gamma(n/2) \geq C\{n/(2e)\}^{n/2-1}\sqrt{n}$ with $C = e\sqrt{\pi}$. Hence, $(1-\xi_n) \leq C/\sqrt{n}$ for some absolute constant $C > 0$.

Proof of Lemma 6.5

Let $s = n/2$, $T = \sum_{j=1}^{n} q_j x_j + q_0$ and $q_j' = (q_j + q_0)$. Then, the multiple integral in Lemma 6.5 equals

$$\int_{\sum_{j=1}^n x_j \leq 1} \frac{1}{T^{s-1}} \prod_{j=1}^{n} x_j^{\alpha_j - 1} dx = \int_{\sum_{j=1}^n x_j \leq 1} \frac{1}{T^{s-1}} \prod_{j=1}^{n} \left(\frac{q_j' x_j}{T}\right)^{\alpha_j - 1} \prod_{j=1}^{n} \left(\frac{T}{q_j'}\right)^{\alpha_j - 1} dx$$
$$= \prod_{j=1}^{n} \left(\frac{1}{q_j'}\right)^{\alpha_j - 1} \int_{\sum_{j=1}^n x_j \leq 1} T^{1-n} \prod_{j=1}^{n} \left(\frac{q_j' x_j}{T}\right)^{\alpha_j - 1} dx \tag{A.1}$$

Now, we make a change of variable from x to z, with $z_j = q'_j x_j / T$ for $j = 1 \ldots, n$. Clearly, z also belongs to the simplex $\Delta^{(n-1)}$. Moreover, letting $z_{n+1} = 1 - \sum_{j=1}^n z_j$, one has $z_{n+1} = q_0 x_{n+1}/T$, where $x_{n+1} = 1 - \sum_{j=1}^n x_j$. Thus, by composition rule,

$$T = \frac{x_1}{\frac{z_1}{q'_1}} = \cdots = \frac{x_n}{z_n/q_n} = \frac{x_{n+1}}{z_n/q_0} = \frac{1}{z_1/q'_1 + \cdots + z_n/q'_n + z_{n+1}/q_0} \tag{A.2}$$

Let $J = \left(\frac{\partial x_j}{\partial z_l}\right)_{jl}$ be the Jacobian of the transformation and $H = \left(\frac{\partial x_j}{\partial z_l}\right)_{jl} = J^{-1}$. Then,

$$H_{jl} = \begin{cases} \frac{q'_j}{T^2}(T - q_j X_j) \text{ if } l = j \\ \frac{-q'_j X_j}{T^2} q_l \text{ if } l \neq j \end{cases}$$

Clearly, $|H| = |H_1| \prod_{j=1}^n \frac{q'_j}{T^2}$ with $H_1 = T\mathrm{I}_n - xq^{\mathrm{T}}$, where $q = (q_1, \ldots, q_n)^{\mathrm{T}}$ and $|A|$ denotes the determinant of a square matrix A. Using a standard result for determinants of rank one perturbations, one has $|H_1| = T^n \left|\mathrm{I}_n - \frac{1}{T} x q^{\mathrm{T}}\right| = T^n\left(1 - \frac{q^{\mathrm{T}} x}{T}\right) = q_0 T^{n-1}$, implying $|H| = (q_0 T^{n-1}) \prod_{j=1}^n \frac{q'_j}{T^2} = \frac{q_0}{T^{n+1}} \prod_{j=1}^n q'_j$. Hence the Jacobian of the transformation is

$$|J| = \frac{T^{n+1}}{r \prod_{j=1}^n q'_j},$$

so that the change of variable in (A.1) results in

$$\begin{aligned} &\prod_{j=1}^n \left(\frac{1}{q'_j}\right)^{\alpha_j - 1} \frac{1}{q_0 \prod_{j=1}^n q'_j} \int_{\sum_{j=1}^n z_j \leq 1} \left\{ \prod_{j=1}^n z_j^{\alpha_j - 1} \right\} T^2 dz \\ &= q_0 \prod_{j=1}^n \left(\frac{1}{q'_j}\right)^{\alpha_j} \int_{\sum_{j=1}^n z_j \leq 1} \left\{ \prod_{j=1}^n z_j^{\alpha_j - 1} \right\} \frac{T^2}{q_0^2} dz \\ &= q_0 \prod_{j=1}^n \left(\frac{1}{q'_j}\right)^{\alpha_j} \int_{\sum_{j=1}^n z_j \leq 1} \frac{1}{(\nu_1 z_1 + \cdots + \nu_n z_n + z_{n+1})^2} \left\{ \prod_{j=1}^n z_j^{\alpha_j - 1} \right\} dz \end{aligned} \tag{A.3}$$

where $v_j = \frac{q_0}{q_j+q_0} = \frac{q_0}{q_j'}$. Now, the expression in (A.3) clearly equals

$$q_0 \prod_{j=1}^{n} \left(\frac{1}{q_j'}\right)^{\alpha_j} \frac{\prod_{j=1}^{n} \Gamma(\alpha_j)}{\Gamma(s+1)} \mathbb{E}\Big\{\nu_1 Z_1 + \cdots + \nu_n Z_n + Z_{n+1}\Big\}^{-2}, \tag{A.4}$$

where $(Z_1, \ldots, Z_n) \sim \text{Dir}(\alpha_1, \ldots, \alpha_n, 1)$. A profound result in [7] shows that expectations of functions of Dirichlet random vectors as above can be reduced to the expectation of a functional of univariate Beta random variable:

Dickey's formula [7]: Let $(Z_1, \cdots, Z_n) \sim \text{Dir}(\beta_1, \cdots, \beta_n, \beta_{n+1})$ and $Z_{n+1} = 1 - \sum_{i=1}^{n} Z_i$. Suppose $a < \sum_{j=1}^{n+1} \beta_j$. Then, for $\nu_j > 0$,

$$\mathbb{E}\Big[\sum_{j=1}^{n+1} \nu_j Z_j\Big]^{-a} = \mathbb{E} \prod_{j=1}^{n+1} \frac{1}{\{\nu_j + X(1-\nu_j)\}^{\alpha_j}},$$

where $X \sim \text{Beta}(b, a)$ with $b = \sum_{j=1}^{n+1} \beta_j - a$.

Applying Dickey's formula with $\beta_j = \alpha_j = 1/2$ for $j = 1, \ldots, n$, $\beta_{n+1} = 1$ and $a = 2$ (so that $b = \frac{n}{2} + 1 - 2 = \frac{n}{2} - 1$), (A.4) reduces to

$$q_0 \prod_{j=1}^{n} \left(\frac{1}{q_j'}\right)^{\alpha_j} \frac{\prod_{j=1}^{n} \Gamma(\alpha_j)}{\Gamma(s+1)} \mathbb{E} \prod_{j=1}^{n} \frac{1}{\big\{\nu_j + X(1-\nu_j)\big\}^{\alpha_j}} \tag{A.5}$$

where $X \sim \text{Beta}(b, a)$ with density $f(x) = (n/2)(n/2-1)x^{n/2-2}(1-x)$ for $x \in (0,1)$. Hence, (A.5) finally reduces to

$$q_0 \left(\frac{n}{2} - 1\right) \frac{\Gamma(1/2)^n}{\Gamma(n/2)} \int_0^1 \frac{x^{n/2-2}(1-x)}{\prod_{j=1}^{n} (q_j x + q_0)^{\alpha_j}} dx$$

□

Proof of Lemma 6.6

A standard inequality (see, for example, Formula 7.1.13 in [1]) states

$$\frac{2}{x+\sqrt{x^2+2}} \leq \sqrt{\pi}e^x \text{erfc}(\sqrt{x}) \leq \frac{2}{x+\sqrt{x+4/\pi}} \tag{A.6}$$

In view of (A.6), to prove (36) it is enough to show that $2\sqrt{x+1/\pi} \leq \sqrt{x}+\sqrt{x+4/\pi}$, which follows since:

$$(\sqrt{x}+\sqrt{x+4/\pi})^2 - 4(x+1/\pi) = 2x + 2\sqrt{x}\sqrt{x+4/\pi} - 4x \geq 0.$$

To show (37), we use the lower bound for the complementary error function in (A.6). First, we will show that for any $\delta > 0$, $x+\sqrt{x^2+2} \leq 2x^{1+\delta}$ if $x \geq 2$. Noting that if $x \geq 2$

$$x^{2+2\delta} - x^2 = x^2(x^{2\delta}-1) = x^2(x-1)(1+x+\cdots+x^{2\delta}) \geq 2.$$

Hence $\sqrt{x^2+2} \leq x^{1+\delta}$ if $x \geq 2$, showing that $x+\sqrt{x^2+2} \leq 2x^{1+\delta}$ if $x \geq 2$. Thus, we have, for $x \geq 2$ and any $\delta > 0$,

$$\sqrt{\pi}e^x \text{erfc}(\sqrt{x}) \geq \left(\frac{1}{\sqrt{x}}\right)^{1+\delta}.$$

□

REFERENCES

[1] M. Abramowitz and I.A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Dover publications, 1965.

[2] A. Armagan, D. Dunson, and J. Lee. Generalized double pareto shrinkage. *Arxiv preprint arxiv:1104.0861*, 2011.

[3] D. Bontemps. Bernstein–von mises theorems for gaussian regression with increasing number of regressors. *The Annals of Statistics*, 39(5):2557–2584, 2011.

[4] C.M. Carvalho, N.G. Polson, and J.G. Scott. Handling sparsity via the horseshoe. *Journal of Machine Learning Research W&CP*, 5(73-80), 2009.

[5] C.M. Carvalho, N.G. Polson, and J.G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[6] I. Castillo and A. van der Vaart. Needles and straws in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069–2101, 2012.

[7] J.M. Dickey. Three multidimensional-integral identities with bayesian applications. *The Annals of Mathematical Statistics*, 39(5):1615–1628, 1968.

[8] D.L. Donoho, I.M. Johnstone, J.C. Hoch, and A.S. Stern. Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 41–81, 1992.

[9] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[10] A. Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006.

[11] S. Ghosal. Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli*, 5(2):315–331, 1999.

[12] IS Gradshteyn and IM Ryzhik. Corrected and enlarged edition. *Tables of Integrals, Series and ProductsAcademic Press, New York*, 1980.

[13] J.E. Griffin and P.J. Brown. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.

[14] C. Hans. Elastic net regression modeling with the orthant normal prior. *Journal of the American Statistical Association*, 106(496):1383–1393, 2011.

[15] I.M. Johnstone and B.W. Silverman. Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649, 2004.

[16] I.M. Johnstone and B.W. Silverman. Ebayesthresh: R and s-plus programs for empirical bayes thresholding. *J. Statist. Soft*, 12:1–38, 2005.

[17] W. Kruijer, J. Rousseau, and A. van der Vaart. Adaptive bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257, 2010.

[18] C. Leng. Variable selection and coefficient estimation via regularized rank regression. *Statistica Sinica*, 20(1):167, 2010.

[19] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, 2009.

[20] S. Negahban, P. Ravikumar, M.J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *arXiv preprint arXiv:1010.2731*, 2010.

[21] T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[22] D. Pati, A. Bhattacharya, N.S. Pillai, and D.B. Dunson. Posterior contraction in sparse bayesian factor models for massive covariance matrices. *arXiv preprint arXiv:1206.3627*, 2012.

[23] N.G. Polson and J.G. Scott. Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian Statistics 9 (J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West, eds.)*, pages 501–538. Oxford University Press, New York, 2010.

[24] N.G. Polson and J.G. Scott. On the half-cauchy prior for a global scale parameter. *arXiv preprint arXiv:1104.4937*, 2011.

[25] G. Raskutti, M.J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over l_q balls. *Information Theory, IEEE Transactions on*, 57(10):6976–6994, 2011.

[26] J.G. Scott and J.O. Berger. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587–2619, 2010.

[27] N. Strawn, A. Armagan, R. Saab, L. Carin, and D. Dunson. Finite sample posterior concentration in high-dimensional regression. *arXiv preprint arXiv:1207.4854*, 2012.

[28] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[29] S.A. Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

[30] AW van der Vaart and JH van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. *IMS Collections*, 3:200–222, 2008.

[31] H. Wang and C. Leng. Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048, 2007.

[32] M. West. On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648, 1987.

[33] C.H. Zhang and J. Huang. Model selection consistency of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 2008.

[34] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(2):2541–2567, 2006.

[35] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.